T.C.
HACETTEPE ÜNİVERSİTESİ
TIP FAKÜLTESİ
İÇ HASTALIKLARI ANABİLİM DALI

GASTROENTEROLOJİ'YE ÖZGÜ YAPAY ZEKA DİL MODELİNİN KLİNİK PRATİKTE KULLANIMI

Dr. Ali Berkcan BOZDOĞAN

UZMANLIK TEZİ

Olarak Hazırlanmıştır

ANKARA

2026

T.C.
HACETTEPE ÜNİVERSİTESİ
TIP FAKÜLTESİ
İÇ HASTALIKLARI ANABİLİM DALI

GASTROENTEROLOJİ'YE ÖZGÜ YAPAY ZEKA DİL MODELİNİN KLİNİK PRATİKTE KULLANIMI

Dr. Ali Berkcan BOZDOĞAN

UZMANLIK TEZİ

Olarak Hazırlanmıştır

TEZ DANIŞMANI

Doç. Dr. Muhammed Bahattin DURAK

ANKARA

2026

TEŞEKKÜR

Bu tez çalışmasının her aşamasında değerli bilgi ve deneyimlerini benimle paylaşan, akademik gelişimime katkıda bulunan ve bilimsel bakış açımı şekillendiren saygıdeğer hocam Doç. Dr. Cem Şimşek'e sonsuz şükranlarımı sunarım. Kendisinin sabırlı rehberliği ve yapıcı eleştirileri bu çalışmanın tamamlanmasında en büyük desteklerden biri olmuştur.

Tezimin planlanması ve yürütülmesi sürecinde değerli katkılarını esirgemeyen, bilgi ve tecrübelerinden faydalandığım Doç. Dr. Muhammed Bahattin Durak'a en içten teşekkürlerimi sunarım. Kendisinin gastroenteroloji alanındaki derin bilgisi ve önerileri bu çalışmanın niteliğini artırmıştır.

Tıp eğitimim ve uzmanlık sürecim boyunca akademik ve mesleki gelişimime katkı sağlayan tüm hocalarıma, çalışma arkadaşlarıma ve hastalarıma teşekkürü bir borç bilirim.

Bu zorlu süreçte yanımda olan, sabır ve anlayışıyla beni destekleyen, her zaman inancını ve sevgisini hissettiren sevgili eşim Fatma Büşra Bozdoğan'a en derin minnettarlığımı ifade ederim. Onun varlığı, bu çalışmanın tamamlanmasında en büyük motivasyon kaynağım olmuştur.

Son olarak, bugünlere gelmemde emeği geçen, eğitimime verdikleri destekle her zaman arkamda duran değerli aileme sonsuz teşekkürlerimi sunarım.

ÖZET

Bozdoğan Ali B.: Gastroenteroloji'ye Özgü Yapay Zeka Dil Modelinin Klinik Pratikte Kullanımı; Hacettepe Üniversitesi Tıp Fakültesi (HÜTF) İç Hastalıkları Anabilim Dalı Uzmanlık Tezi; Ankara, 2026.

Amaç: Bu çalışma, gastroenteroloji alanına özgü olarak geliştirilen retrieval-augmented generation (RAG) tabanlı bir yapay zeka dil modeli olan GastroGPT'nin klinik karar destek kapasitesini değerlendirmeyi ve genel amaçlı büyük dil modelleri ile karşılaştırmalı analizini yapmayı amaçlamaktadır.

Gereç ve Yöntem: Retrospektif, çok merkezli bu karşılaştırmalı model değerlendirme çalışmasında, Hacettepe Üniversitesi Tıp Fakültesi Gastroenteroloji Bilim Dalı arşivinden seçilen 200 standardize klinik vaka kullanılmıştır. Vakalar; genel gastroenteroloji (%50), hepatoloji (%26), pankreatik hastalıklar (%9), inflamatuvar barsak hastalıkları (%6), gastrointestinal onkoloji (%5) ve endoskopi (%4) alt uzmanlık alanlarını temsil edecek şekilde stratifiye edilmiştir. Çalışmada dört GastroGPT varyantı (Deeplake, Faiss, Chroma ve 16K) ile dört genel amaçlı model (GPT-4, Claude, Bard/Gemini, You.com) olmak üzere toplam sekiz yapay zeka dil modeli değerlendirilmiştir. Model yanıtları, iki bağımsız uzman gastroenterolog tarafından körleştirilmiş olarak; tanısal doğruluk, önerilen tetkiklerin uygunluğu, tedavi önerilerinin kılavuzlara uygunluğu, hasta yönetimi stratejilerinin uygulanabilirliği ve genel performans olmak üzere beş kriter üzerinden 5'li Likert ölçeği ile puanlanmıştır. Ek olarak halüsinasyon oranları, kaynak atfı kalitesi ve yanıt tutarlılığı değerlendirilmiştir.

Bulgular: GastroGPT-Deeplake, genel performansta en yüksek puanı elde etmiştir (53.8±5.9; maksimum 60 puan) ve GPT-4 (50.1±6.8, p=0.028) ile Claude (49.3±7.2, p=0.012) modellerini istatistiksel olarak anlamlı düzeyde geride bırakmıştır. GastroGPT varyantları, halüsinasyon oranları (%4.2-6.8'e karşı %8.3-12.1, p=0.018) ve kaynak atfı kalitesi (%62.8-78.3'e karşı %23.5-34.7) açısından genel amaçlı modellere göre belirgin üstünlük göstermiştir. Vektör veri tabanı karşılaştırmasında Deeplake, Faiss ve Chroma'ya göre anlamlı düzeyde üstün performans sergilemiştir (p<0.001). GastroGPT'nin beş değerlendirme kriterindeki

ortalama puanları 4.00 ile 4.21 arasında değişmiş olup, en yüksek performans kılavuz uyumu kriterinde (4.21±0.63) elde edilmiştir. Vaka karmaşıklığı tanısal doğruluk üzerinde anlamlı bir etkiye sahip bulunmuş (F=7.361, p=0.0008); yüksek karmaşıklıktaki vakalarda tanısal doğruluk puanı (3.81±0.83) düşük karmaşıklıktaki vakalara (4.43±0.70) kıyasla anlamlı düzeyde düşük saptanmıştır (Cohen's d=0.803). Hastalık görülme sıklığı ve alt uzmanlık alanları açısından anlamlı performans farklılığı gözlenmemiştir. Değerlendiriciler arası güvenilirlik yüksek düzeyde bulunmuştur (ICC>0.99, Pearson r=0.814, Cronbach α=0.802). ROC analizinde genel performans kriteri için AUC değeri 0.908 olarak hesaplanmıştır.

Sonuç: GastroGPT-Deeplake, alan-spesifik RAG entegrasyonu sayesinde genel amaçlı büyük dil modellerine göre hem genel performans hem de güvenlik parametreleri açısından üstünlük göstermiştir. RAG teknolojisinin halüsinasyon riskini azaltma ve kaynak izlenebilirliği sağlama konusundaki avantajları, tıbbi yapay zeka uygulamalarının güvenliği için kritik öneme sahiptir. Bulgularımız, gastroenteroloji pratiğinde yapay zeka destekli klinik karar destek sistemlerinin geliştirilmesinde alan-spesifik yaklaşımların genel amaçlı modellere tercih edilmesi gerektiğini desteklemektedir.

Anahtar Kelimeler: Yapay zeka, büyük dil modeli, gastroenteroloji, GastroGPT, retrieval-augmented generation, klinik karar destek sistemi, halüsinasyon, vektör veri tabanı

ABSTRACT

Bozdoğan Ali B.: Clinical Use of a Gastroenterology-Specific Artificial Intelligence Language Model; Hacettepe University Adult Hospital Residency Thesis, Department of Internal Medicine, Hacettepe University Faculty of Medicine, Ankara, 2026.

Objective: This study aimed to evaluate the clinical decision support capacity of GastroGPT, a retrieval-augmented generation (RAG)-based artificial intelligence language model developed specifically for gastroenterology, and to perform a comparative analysis with general-purpose large language models.

Materials and Methods: In this retrospective, multicenter comparative model evaluation study, 200 standardized clinical cases selected from the archives of Hacettepe University Faculty of Medicine, Department of Gastroenterology were utilized. Cases were stratified to represent the following subspecialty areas: general gastroenterology (50%), hepatology (26%), pancreatic diseases (9%), inflammatory bowel diseases (6%), gastrointestinal oncology (5%), and endoscopy (4%). A total of eight AI language models were evaluated, comprising four GastroGPT variants (Deeplake, Faiss, Chroma, and 16K) and four general-purpose models (GPT-4, Claude, Bard/Gemini, You.com). Model responses were scored by two independent expert gastroenterologists in a blinded fashion using a 5-point Likert scale across five criteria: diagnostic accuracy, appropriateness of recommended investigations, guideline concordance of treatment recommendations, applicability of patient management strategies, and overall performance. Additionally, hallucination rates, source attribution quality, and response consistency were assessed.

Results: GastroGPT-Deeplake achieved the highest overall performance score (53.8±5.9; maximum 60 points), significantly outperforming GPT-4 (50.1±6.8, $p=0.028$) and Claude (49.3±7.2, $p=0.012$). GastroGPT variants demonstrated marked superiority over general-purpose models in hallucination rates (4.2–6.8% vs. 8.3–12.1%, $p=0.018$) and source attribution quality (62.8–78.3% vs. 23.5–34.7%). In the vector database comparison, Deeplake significantly outperformed Faiss and Chroma ($p<0.001$). GastroGPT's mean scores across the five evaluation criteria ranged from

4.00 to 4.21, with the highest performance observed in guideline concordance (4.21±0.63). Case complexity had a significant effect on diagnostic accuracy (F=7.361, p=0.0008); diagnostic accuracy scores in high-complexity cases (3.81±0.83) were significantly lower compared to low-complexity cases (4.43±0.70; Cohen's d=0.803). No significant performance differences were observed across disease prevalence categories or subspecialty areas. Inter-rater reliability was excellent (ICC>0.99, Pearson r=0.814, Cronbach's α=0.802). ROC analysis yielded an AUC of 0.908 for the overall performance criterion.

Conclusion: GastroGPT-Deeplake demonstrated superiority over general-purpose large language models in both overall performance and safety parameters through domain-specific RAG integration. The advantages of RAG technology in reducing hallucination risk and ensuring source traceability are of critical importance for the safety of medical AI applications. Our findings support the preferential use of domain-specific approaches over general-purpose models in developing AI-assisted clinical decision support systems for gastroenterology practice.

Keywords: Artificial intelligence, large language model, gastroenterology, GastroGPT, retrieval-augmented generation, clinical decision support system, hallucination, vector database

İÇİNDEKİLER

SİMGELER VE KISALTMALAR

AGA : American Gastroenterological Association (Amerikan Gastroenteroloji Derneği)

ACG : American College of Gastroenterology (Amerikan Gastroenteroloji Koleji)

AI : Artificial Intelligence (Yapay Zeka)

ANOVA : Analysis of Variance (Varyans Analizi)

API : Application Programming Interface (Uygulama Programlama Arayüzü)

ASGE : American Society for Gastrointestinal Endoscopy (Amerikan Gastrointestinal Endoskopi Derneği)

AUC : Area Under the Curve (Eğri Altında Kalan Alan)

BISAP : Bedside Index for Severity in Acute Pancreatitis (Akut Pankreatitte Şiddet İndeksi)

CADe : Computer-Aided Detection (Bilgisayar Destekli Tespit)

CADx : Computer-Aided Diagnosis (Bilgisayar Destekli Tanı)

CI : Confidence Interval (Güven Aralığı)

Cronbach α : İç tutarlılık katsayısı

ECCO : European Crohn's and Colitis Organisation (Avrupa Crohn ve Kolit Organizasyonu)

EASL : European Association for the Study of the Liver (Avrupa Karaciğer Araştırmaları Derneği)

Faiss : Facebook AI Similarity Search (Vektör Benzerlik Arama Kütüphanesi)

F : F testi istatistiği

GA : Güven Aralığı

GCP : Good Clinical Practice (İyi Klinik Uygulamalar)

GI : Gastrointestinal

GPT : Generative Pre-trained Transformer

GastroGPT : Gastroenterolojiye Özgü Yapay Zeka Dil Modeli

ICC : Intraclass Correlation Coefficient (Sınıf İçi Korelasyon Katsayısı)

IBH : İnflamatuvar Barsak Hastalıkları

IQR : Interquartile Range (Çeyrekler Arası Aralık)

LLM : Large Language Model (Büyük Dil Modeli)

MAR : Missing at Random

MCAR : Missing Completely at Random

MNAR : Missing Not at Random

MELD : Model for End-Stage Liver Disease

NPV : Negative Predictive Value (Negatif Prediktif Değer)

OR : Odds Ratio

PPV : Positive Predictive Value (Pozitif Prediktif Değer)

RAG : Retrieval-Augmented Generation (Geri Getirim Destekli Metin Üretimi)

ROC : Receiver Operating Characteristic

SD : Standard Deviation (Standart Sapma)

SOAP : Subjective, Objective, Assessment, Plan (Klinik Değerlendirme Formatı)

TNM : Tumor–Node–Metastasis (Tümör–Lenf Nodu–Metastaz Evreleme Sistemi)

VIF : Variance Inflation Factor (Varyans Şişirme Faktörü)

YZ : Yapay Zeka

TABLOLAR

1. GİRİŞ VE AMAÇ

Yapay zeka (YZ) teknolojileri, son on yılda tıp pratiğinin hemen her alanında köklü dönüşümlere öncülük etmiştir(1, 2). Görüntü tanıma algoritmalarının radyoloji ve patoloji alanlarındaki başarısı, klinik karar destek sistemlerinin acil servislerde ve yoğun bakım ünitelerindeki yaygınlaşması ve doğal dil işleme tabanlı uygulamaların elektronik sağlık kayıtlarından bilgi çıkarımında elde ettiği kazanımlar, sağlık hizmetlerinin sunumunda paradigma değişikliklerine yol açmıştır (3). Bu teknolojik devrim, tanısal doğruluğun artırılması, tedavi süreçlerinin optimize edilmesi ve sağlık hizmetlerinin maliyetlerinin düşürülmesi gibi çok boyutlu kazanımlar vadetmektedir (4). Özellikle derin öğrenme algoritmalarının gelişimi, tıbbi görüntülerin yorumlanmasından genom analizine, ilaç keşfinden hasta izleme sistemlerine kadar geniş bir yelpazede çığır açan uygulamalara zemin hazırlamıştır (5) (6).

Gastroenteroloji disiplini, hem görüntüleme yoğunluğu hem de klinik karar verme karmaşıklığı açısından bu teknolojik dönüşümden en fazla etkilenen alanlardan biri olarak öne çıkmaktadır (7). Endoskopik prosedürlerin yüksek görüntü çıktısı, hepatolojik değerlendirmelerin çok değişkenli doğası ve inflamatuvar bağırsak hastalıklarının uzun dönemli takip gereksinimleri, gastroenterolojide YZ uygulamalarının geliştirilmesi için verimli bir zemin oluşturmuştur (8). Gastrointestinal sistem hastalıkları, dünya genelinde önemli bir morbidite ve mortalite nedeni olmaya devam etmekte olup, kolorektal kanser tek başına kanser ilişkili ölümlerin önde gelen nedenlerinden birini oluşturmaktadır (9, 10). Bu epidemiyolojik gerçeklik, gastroenteroloji alanında tanı ve tedavi süreçlerinin optimize edilmesinin önemini vurgulamaktadır.

Büyük dil modelleri (Large Language Models, LLM), yapay zekanın metin anlama ve üretme kapasitesinde devrim niteliğinde ilerlemeler kaydeden bir alt dalı temsil etmektedir(11) (12). GPT (Generative Pre-trained Transformer) mimarisi üzerine inşa edilen bu modeller, milyarlarca parametre içeren derin sinir ağları aracılığıyla insan dilinin karmaşık yapısını öğrenmekte ve bağlama uygun, tutarlı metinler üretebilmektedir (13). Transformer mimarisinin dikkat mekanizması (attention mechanism), modellerin uzun bağlamlarda bile anlamlı ilişkiler kurmasına

olanak tanımaktadır. Bu teknolojik atılım, 2017 yılında Vaswani ve arkadaşları tarafından sunulan çığır açıcı çalışmanın ardından hızla gelişmiştir. Tıp alanında LLM'lerin potansiyel uygulamaları arasında klinik dokümantasyon, hasta eğitimi, literatür sentezi, tanısal muhakeme desteği ve tedavi planlaması yer almaktadır (14).OpenAI'nin GPT-4 modeli, Amerika Birleşik Devletleri Tıp Lisans Sınavı (USMLE) simülasyonlarında geçme eşiğini aşan performans sergileyerek tıbbi bilgi kapasitesini kanıtlamıştır (15). Benzer şekilde Google'ın Med-PaLM ve Med-PaLM 2 modelleri, tıbbi soru-cevap görevlerinde uzman düzeyine yaklaşan sonuçlar elde etmiştir. Özellikle retrieval-augmented generation (RAG) teknolojisinin entegrasyonu ile LLM'ler, güncel tıbbi literatür ve kurumsal bilgi tabanlarına erişerek daha güvenilir çıktılar üretme kapasitesi kazanmıştır (16).

Kolonoskopi alanında bilgisayar destekli tespit (computer-aided detection, CADe) ve bilgisayar destekli tanı (computer-aided diagnosis, CADx) sistemleri, yapay zekanın gastroenterolojideki en olgun uygulama alanını oluşturmaktadır (17). Randomize kontrollü çalışmaların sayısı 40'ı aşmış olup, gastroenteroloji bu açıdan YZ araştırmalarında öncü disiplinler arasında anılmaktadır (18). CADe sistemlerinin adenom saptama oranını (adenoma detection rate, ADR) anlamlı düzeyde artırdığı ortaya konmuştur (19). Meta-analizler, CADe kullanımının ADR'yi yaklaşık %30 oranında artırdığını göstermiştir (20). Gastroenterolojiye özgü LLM geliştirme çalışmaları, alan-spesifik ince ayar (domain-specific fine-tuning) ve RAG entegrasyonu üzerine yoğunlaşmaktadır. RAG teknolojisi, modelin çıktı üretirken güncel ve güvenilir kaynaklara başvurmasını sağlayarak halüsinasyon riskini azaltmakta ve kaynak atfı imkânı sunmaktadır. Vektör veri tabanları aracılığıyla semantik benzerlik araması yapılarak, kullanıcı sorgusuna en uygun doküman parçaları modele bağlam olarak sunulmaktadır (21). Amerikan Gastroenteroloji Derneği (AGA), Avrupa Karaciğer Araştırmaları Derneği (EASL), Amerikan Gastrointestinal Endoskopi Derneği (ASGE) ve Avrupa Crohn ve Kolit Örgütü (ECCO) rehberleri, RAG sistemleri için ideal bilgi kaynakları oluşturmaktadır (22)

Vektör veri tabanları ve embedding teknolojileri, RAG sistemlerinin performansını belirleyen kritik bileşenlerdir (23). Deeplake, Faiss ve Chroma gibi farklı vektör veri tabanı çözümleri, sorgu-doküman eşleştirmesinde farklı algoritmalar

kullanmaktadır. Faiss (Facebook AI Similarity Search), büyük ölçekli vektör aramaları için optimize edilmiştir. Chroma basit API yapısı ile, Deeplake ise versiyon kontrolü özellikleriyle öne çıkmaktadır. (24) LLM'lerin klinik karar desteğinde kullanımı, güvenlik ve etik boyutlarıyla birlikte değerlendirilmelidir Halüsinasyon, aşırı güven ve belirsizlik kalibrasyonu sorunları, bu sistemlerin klinik ortamda güvenli kullanımını tehdit eden başlıca riskler arasındadır (25). Klinisyen-YZ etkileşiminde insan-onay döngüsü ilkesinin benimsenmesi, güvenli uygulama için temel gereklilikler olarak kabul edilmektedir (26).

Bu tez çalışması, gastroenteroloji alanına özgü bir yapay zeka dil modelinin (GastroGPT) klinik pratikte kullanımını değerlendirmeyi amaçlamaktadır. Çalışmanın birincil amaçları: (1) Gastroenterolojiye özgü GastroGPT modelinin farklı vektör veritabanı konfigürasyonlarındaki performansını değerlendirmek, (2) GastroGPT'yi GPT-4, Claude ve Bard gibi genel amaçlı LLM'lerle karşılaştırmak, (3) Model performansının vaka karmaşıklığı, hastalık sıklığı ve alt uzmanlık alanına göre değişimini analiz etmektir.

2. MATERYAL VE METOT

2.1. Çalışma Tasarımı

Bu çalışma, gastroenteroloji alanına özgü bir yapay zeka dil modelinin (GastroGPT) klinik karar destek kapasitesini değerlendirmek amacıyla tasarlanmış retrospektif, çok merkezli bir karşılaştırmalı model değerlendirme çalışmasıdır. Çalışma protokolü, Hacettepe Üniversitesi Girişimsel Olmayan Klinik Araştırmalar Etik Kurulu tarafından onaylanmıştır (Karar No: SBA 24/979, Tarih: 17.09.2024). Çalışma, Helsinki Deklarasyonu ilkelerine ve İyi Klinik Uygulamalar (GCP) kılavuzlarına uygun olarak yürütülmüştür. Çalışmanın birincil sonlanım noktası, modellerin SOAP notu formatındaki klinik vaka değerlendirmelerinde elde ettikleri toplam performans puanı olarak belirlenmiştir. İkincil sonlanım noktaları arasında alt uzmanlık alanlarına göre performans, vaka karmaşıklığına göre performans, halüsinasyon oranları, yanıt tutarlılığı ve kaynak atfı kalitesi yer almaktadır.

2.2. Vaka Seti ve Seçim Kriterleri

Çalışmaya dahil edilen vaka seti, gastroenteroloji pratiğinin farklı alt alanlarını temsil eden 200 standardize klinik vakadan oluşmaktadır. Vakalar, Hacettepe Üniversitesi Tıp Fakültesi Gastroenteroloji Bilim Dalı arşivinden retrospektif olarak seçilmiş gerçek klinik senaryolardan adapte edilmiştir. Her vaka, en az iki uzman gastroenterolog (deneyim süresi ≥10 yıl) tarafından bağımsız olarak gözden geçirilmiş ve klinik doğruluğu onaylanmıştır. Vaka seçim kriterleri şu şekilde belirlenmiştir: (1) Gastroenteroloji pratiğinde karşılaşılan tipik klinik senaryoları temsil etme, (2) Farklı karmaşıklık düzeylerini içerme, (3) Çeşitli alt uzmanlık alanlarını kapsama, (4) Tanısal ve terapötik karar verme süreçlerini test etmeye uygun olma. Dışlama kriterleri olarak; eksik klinik bilgi içeren vakalar, çoklu organ tutulumu nedeniyle gastroenteroloji dışı uzmanlık gerektiren vakalar ve nadir sendromların atipik prezantasyonları belirlenmiştir.

Her vaka, yapılandırılmış bir formatta sunulmuştur. Bu format şu bileşenleri içermektedir: (a) Demografik bilgiler (yaş, cinsiyet, meslek, yaşam koşulları), (b) Başvuru şikayeti ve süresi, (c) Anamnez (hastalık öyküsü, özgeçmiş, soy geçmiş,

alışkanlıklar, kullanılan ilaçlar, alerji öyküsü), (d) Sistem sorgulaması, (e) Fizik muayene bulguları, (f) Varsa laboratuvar ve görüntüleme sonuçları. Vakalar, hasta mahremiyetini korumak amacıyla herhangi bir kişisel tanımlayıcı bilgi içermemektedir. Demografik özellikler açısından vakalar 18-76 yaş aralığında (ortalama: 46.4±15.2 yıl, medyan: 47 yıl, IQR: 34-58 yıl) hastalardan oluşmaktadır. Cinsiyet dağılımı erkek (n=105, %52.5) ve kadın (n=95, %47.5) olarak dengeli tutulmuştur.

Vaka dağılımı alt uzmanlık alanlarına göre stratifiye edilmiştir: Genel gastroenteroloji (n=100, %50), hepatoloji (n=52, %26), pankreatik hastalıklar (n=18, %9), inflamatuvar bağırsak hastalıkları (n=12, %6), gastrointestinal onkoloji (n=10, %5) ve endoskopi (n=8, %4). Bu dağılım, Türkiye'deki gastroenteroloji kliniklerinde karşılaşılan vaka sıklıklarını ve uzmanlık eğitimi müfredatını yansıtacak şekilde planlanmıştır. Hastalık sıklığına göre sınıflandırma Orphanet nadir hastalık veritabanı ve epidemiyolojik literatür verilerine dayanmaktadır: Sık görülen (prevalans >1/1000, n=80, %40), nadir olmayan (prevalans 1/10000-1/1000, n=83, %41.5) ve nadir (prevalans <1/10000, n=37, %18.5).

Vakalar karmaşıklık düzeylerine göre üç kategoride sınıflandırılmıştır. Düşük karmaşıklık (n=29, %14.5): Tek bir olası tanı, standart tetkik algoritması, birinci basamak tedavi yeterli. Orta karmaşıklık (n=115, %57.5): 2-4 olası tanı, çoklu tetkik gerekliliği, tedavi bireyselleştirmesi gerekli. Yüksek karmaşıklık (n=56, %28): 5 veya daha fazla olası tanı, ileri tetkikler ve multidisipliner yaklaşım gerekli, nadir hastalık olasılığı yüksek. Karmaşıklık sınıflandırması için modifiye Delphi yöntemi kullanılmış, iki bağımsız değerlendirici tarafından yapılan sınıflandırmada uyuşmazlık durumlarında üçüncü bir hakemle konsensüs sağlanmıştır.

2.3. Değerlendirilen Modeller

Çalışmada toplam 8 farklı yapay zeka dil modeli değerlendirilmiştir. Bu modeller iki ana kategoride gruplandırılmıştır: Birinci kategori gastroenterolojiye özgü GastroGPT varyantlarını içermektedir: (1) GastroGPT-Deeplake: Deeplake vektör veritabanı entegreli versiyon, (2) GastroGPT-Faiss: Facebook AI Similarity Search (Faiss) vektör veritabanı entegreli versiyon, (3) GastroGPT-Chroma: Chroma

vektör veritabanı entegreli versiyon, (4) GastroGPT-16K: 16.384 token genişletilmiş bağlam pencereli versiyon. İkinci kategori genel amaçlı büyük dil modellerini içermektedir: (1) GPT-4 (OpenAI, gpt-4-0613 versiyonu), (2) Claude (Anthropic, claude-2 versiyonu), (3) Bard/Gemini (Google, Bard Pro versiyonu), (4) You.com (You.com AI asistan). Embedding işlemi için OpenAI'nin text-embedding-ada-002 modeli kullanılmıştır (1536 boyutlu vektörler). Chunk boyutu 1000 token, overlap 200 token olarak ayarlanmıştır. Retrieval aşamasında her sorgu için en benzer 5 doküman parçası (k=5) modele bağlam olarak sunulmuştur. Benzerlik ölçütü olarak kosinüs benzerliği kullanılmıştır.

GastroGPT modeli, gastroenteroloji alanına özgü olarak geliştirilmiş retrieval-augmented generation (RAG) tabanlı bir sistemdir. Temel LLM olarak GPT-3.5-turbo (OpenAI) üzerine inşa edilmiştir. Bilgi tabanı şu kaynaklardan oluşmaktadır: American Gastroenterology Association (AGA) klinik uygulama kılavuzları (n=127), European Association for the Study of the Liver (EASL) rehberleri (n=89), American Society for Gastrointestinal Endoscopy (ASGE) kılavuzları (n=156), European Crohn's and Colitis Organisation (ECCO) kılavuzları (n=78), American College of Gastroenterology (ACG) kılavuzları (n=134), UpToDate gastroenteroloji bölümleri (n=263). Toplam bilgi tabanı 847 doküman, 12.456 sayfa ve yaklaşık 4,2 milyon kelime içermektedir.

2.4. Değerlendirme Metodolojisi

Her vaka için tüm modellere aynı standart yönlendirme şablonu sunulmuştur: 'Act as a gastroenterology doctor and provide a SOAP note that includes: summary, additional questions, additional tests, proposed management, follow-up plan, referrals and patient counseling.' Bu şablon, klinik muhakeme sürecinin tüm bileşenlerini kapsayacak şekilde tasarlanmış olup SOAP (Subjective, Objective, Assessment, Plan) notu formatını temel almaktadır. Model yanıtları, iki bağımsız gastroenteroloji uzmanı (Değerlendirici A: 15 yıl deneyim, Değerlendirici B: 12 yıl deneyim) tarafından körleştirilmiş olarak değerlendirilmiştir. Değerlendiriciler, hangi yanıtın hangi modele ait olduğunu bilmemektedir. Her vaka için model yanıtları rastgele sıralanarak sunulmuştur.

Değerlendirme kriterleri ve puanlama şeması şu şekilde belirlenmiştir: (1) Tanısal doğruluk (0-10 puan): Doğru tanıya ulaşma veya doğru tanıyı ayırıcı tanılar arasında önceliklendirme. (2) Ayırıcı tanı kapsamlılığı (0-10 puan): Klinik tabloya uygun ayırıcı tanıların eksiksiz ve mantıklı sıralanması. (3) Ek tetkik önerilerinin uygunluğu (0-10 puan): Tanısal algoritmalara ve güncel kılavuzlara uygun tetkik önerileri. (4) Tedavi planının rehber uyumu (0-10 puan): Güncel klinik uygulama kılavuzlara uygun tedavi önerileri. (5) Hasta danışmanlığı kalitesi (0-10 puan): Hasta eğitimi, yaşam tarzı önerileri ve takip planının uygunluğu. (6) Genel klinik muhakeme kalitesi (0-10 puan): Bütüncül klinik düşünme, mantıksal tutarlılık ve profesyonellik.7

2.5. İstatistiksel Analiz

Ek değerlendirme parametreleri olarak şunlar kaydedilmiştir: Halüsinasyon varlığı (var/yok ve türü), kaynak atfı kalitesi (uygun referans var/yok), yanıt tutarlılığı (aynı vaka için tekrarlanan sorgularda benzerlik), yanıt süresi (saniye cinsinden). Örneklem büyüklüğü hesaplaması, pilot çalışma verilerine dayanarak yapılmıştır. İki model arasında 5 puanlık fark tespit edebilmek için (standart sapma: 8, alfa: 0.05, güç: 0.80), her model için minimum 41 vaka gerektiği hesaplanmıştır. Çoklu karşılaştırmalar ve alt grup analizleri için yeterli güç sağlamak amacıyla toplam 200 vaka dahil edilmiştir.

Tanımlayıcı istatistikler için sürekli değişkenler ortalama ± standart sapma (normal dağılım durumunda) veya medyan ve çeyrekler arası aralık (IQR, normal dağılım dışı durumda) olarak sunulmuştur. Kategorik değişkenler frekans ve yüzde olarak ifade edilmiştir. Normal dağılım varsayımı Shapiro-Wilk testi ile değerlendirilmiştir. Grup karşılaştırmaları için şu testler uygulanmıştır: İki bağımsız grup karşılaştırmasında normal dağılım varsa bağımsız örneklem t-testi, yoksa Mann-Whitney U testi; ikiden fazla bağımsız grup karşılaştırmasında normal dağılım varsa tek yönlü varyans analizi (ANOVA), yoksa Kruskal-Wallis testi kullanılmıştır. ANOVA veya Kruskal-Wallis testinde anlamlı fark saptandığında post-hoc çoklu karşılaştırmalar için Bonferroni düzeltmeli pairwise karşılaştırmalar yapılmıştır. Tekrarlı ölçümler için (aynı vaka üzerinde farklı model karşılaştırmaları) Friedman testi ve Wilcoxon işaretli sıralar testi uygulanmıştır. Kategorik değişkenler arasındaki

ilişki için Pearson ki-kare testi veya Fisher'in kesin testi (beklenen frekans <5 olan hücre sayısı>%20 olduğunda) kullanılmıştır.

Değerlendiriciler arası güvenilirlik için DF Cohen's kappa (κ) katsayısı hesaplanmıştır. Kappa değerleri şu şekilde yorumlanmıştır: κ<0.20 zayıf, 0.21-0.40 düşük, 0.41-0.60 orta, 0.61-0.80 iyi, 0.81-1.00 mükemmel uyum. Sınıf içi korelasyon katsayısı (ICC, two-way random, absolute agreement) sürekli değişkenler için hesaplanmıştır. Korelasyon analizleri için normal dağılım varsayımına göre Pearson veya Spearman korelasyon katsayıları hesaplanmıştır. Çoklu değişkenli analizler için çoklu doğrusal regresyon modeli kurulmuş, model varsayımları (doğrusallık, normallik, homojenlik, multikollinearite) kontrol edilmiştir. Multikollinearite değerlendirmesi için varyans şişirme faktörü (VIF) hesaplanmış, VIF>10 olan değişkenler modelden çıkarılmıştır. Eksik veri analizi için eksiklik mekanizması (MCAR, MAR, MNAR) değerlendirilmiştir. Eksik veri oranı <%5 olan değişkenler için listwise deletion, %5-20 arası için çoklu impütasyon (multiple imputation, m=5) uygulanmıştır. Tüm istatistiksel testlerde iki yönlü $p<0.05$ değeri istatistiksel olarak anlamlı kabul edilmiştir.

3. BULGULAR

3.1. Çalışma Popülasyonu ve Demografik Özellikler

Hacettepe Üniversitesi Tıp Fakültesi Gastroenteroloji polikliniğine 01.01.2015-30.11.2024 tarihleri arasında başvuran hastalardan belirlenen dahil edilme ve dışlama kriterlerini karşılayan toplam 200 vaka değerlendirmeye alındı. Çalışmaya dahil edilen vakaların demografik ve klinik özellikleri Tablo 3.1'de detaylı olarak sunulmaktadır.

Tablo 3.1. Çalışma Popülasyonunun Demografik ve Klinik Özellikleri (n=200)

Değişken	n / Ortalama	% / SD
Yaş (yıl)	46.4	± 15.2
Yaş Aralığı	18-76	-
Cinsiyet		
Erkek	105	52.5%
Kadın	95	47.5%
Vaka Karmaşıklığı		
Düşük	29	14.5%
Orta	115	57.5%
Yüksek	56	28.0%
Hastalığın Sıklığı		
Yaygın	80	40.0%
Nadir Olmayan	83	41.5%
Nadir	37	18.5%
Alt Uzmanlık Alanı		
Genel Gastroenteroloji	100	50.0%
Hepatoloji	52	26.0%
Pankreas	18	9.0%
İBH	12	6.0%
GI Onkoloji	10	5.0%
Endoskopi	8	4.0%

SD: Standart Sapma, GI: Gastrointestinal, İBH: İnflamatuvar Barsak Hastalıkları

Çalışma popülasyonunun yaş ortalaması 46.4 ± 15.2 yıl olarak hesaplanmış olup, yaş dağılımı 18 ile 76 yaş arasında geniş bir yelpazede değişim göstermektedir. Cinsiyet dağılımı incelendiğinde, vakaların %52.5'inin (n=105) erkek, %47.5'inin (n=95) kadın hastalardan oluştuğu tespit edilmiştir.

Vaka karmaşıklığı açısından değerlendirildiğinde, vakaların %14.5'i (n=29) düşük karmaşıklıkta, %57.5'i (n=115) orta karmaşıklıkta ve %28.0'i (n=56) yüksek karmaşıklıkta olarak sınıflandırılmıştır. Bu dağılım, gastroenteroloji pratiğinde karşılaşılan gerçek dünya vaka yelpazesinin temsilci bir örneğini oluşturmaktadır. Orta karmaşıklıktaki vakaların çoğunlukta olması, günlük klinik pratiği yansıtmakta ve modelin en sık karşılaşılan vaka tiplerindeki performansının değerlendirilmesine olanak tanımaktadır. Hastalık görülme sıklığı açısından yapılan kategorize işleminde vakaların %40.0'ı (n=80) yaygın görülen hastalıklar, %41.5'i (n=83) nadir olmayan hastalıklar ve %18.5'i (n=37) nadir hastalıklar grubunda yer almıştır. Alt uzmanlık alanlarına göre dağılım incelendiğinde, en sık karşılaşılan alan Genel Gastroenteroloji (%50.0, n=100) olup, bunu sırasıyla Hepatoloji (%26.0, n=52), Pankreas hastalıkları (%9.0, n=18), İnflamatuvar Barsak Hastalıkları (%6.0, n=12), Gastrointestinal Onkoloji (%5.0, n=10) ve Endoskopi (%4.0, n=8) izlemektedir.

3.2. GastroGPT Genel Performans Değerlendirmesi

GastroGPT'nin klinik performansı, beş ana değerlendirme kriteri üzerinden uzman gastroenterologlar tarafından 5'li Likert ölçeği kullanılarak sistematik bir şekilde değerlendirilmiştir. Değerlendirme kriterleri tanısal doğruluk, önerilen tetkiklerin uygunluğu, tedavi önerilerinin kılavuzlara uygunluğu, hasta yönetimi stratejilerinin uygulanabilirliği ve genel performans olarak belirlenmiştir. Her bir kriter için elde edilen performans puanları Tablo 3.2'de detaylı olarak sunulmaktadır.

Tablo 3.2. GastroGPT Değerlendirme Kriterleri Performans Puanları

Kriter	Ortalama	± SD	Medyan	IQR	%95 GA
Tanısal Doğruluk	4.07	0.76	4.0	4.0-5.0	3.96-4.18
Tetkik Uygunluğu	4.08	0.59	4.0	4.0-5.0	3.99-4.16
Kılavuz Uyumu	4.21	0.63	4.0	4.0-5.0	4.12-4.30
Yönetim Uygulanabilirliği	4.00	0.69	4.0	4.0-4.0	3.90-4.10
Genel Performans	4.18	0.59	4.0	4.0-5.0	4.10-4.26
Toplam Puan	**20.54**	**2.44**	**21.0**	**19.0-22.0**	**20.20-20.88**

SD: Standart Sapma, IQR: Çeyrekler Arası Aralık, GA: Güven Aralığı. Puanlama: 1=Çok yetersiz, 2=Yetersiz, 3=Orta, 4=İyi, 5=Çok iyi. Maksimum toplam puan: 25

Tanısal doğruluk değerlendirmesinde, GastroGPT 4.07 ± 0.76 ortalama puan elde etmiştir (medyan: 4.0, IQR: 4.0-5.0, %95 güven aralığı: 3.96-4.18). Vakaların %80.0'ında (n=160) tanısal doğruluk puanı klinik kabul edilebilirlik eşiği olan 4 ve üzerinde bulunmuştur. Puan dağılımı incelendiğinde, vakaların %27.5'inde (n=55) çok iyi (5 puan), %56.5'inde (n=113) iyi (4 puan), %15.5'inde (n=31) orta (3 puan) ve yalnızca %0.5'inde (n=1) yetersiz (2 puan) değerlendirme yapılmıştır. Önerilen tetkiklerin uygunluğu açısından model 4.08 ± 0.59 ortalama puan almıştır (%95 güven aralığı: 3.99-4.16). Bu kriterde vakaların %87.0'sında (n=174) klinik olarak kabul edilebilir düzeyde performans sergilenmiştir. Vakaların %26.5'inde (n=53) tamamen uygun, %62.0'ında (n=124) uygun ve %11.0'ında (n=22) kısmen uygun olarak değerlendirme yapılmıştır.

Tedavi önerilerinin kılavuzlara uygunluğu kriterinde GastroGPT beş kriter arasında en yüksek performansını sergilemiş ve 4.21 ± 0.63 ortalama puan elde etmiştir (%95 güven aralığı: 4.12-4.30). Vakaların %88.5'inde (n=177) tedavi önerileri güncel gastroenteroloji kılavuzlarıyla uyumlu bulunmuştur. Puan dağılımı değerlendirildiğinde, vakaların %33.5'inde (n=67) tamamen uyumlu, %55.5'inde (n=111) uyumlu ve %11.0'ında (n=22) kısmen uyumlu olarak değerlendirme yapıldığı görülmüştür.

Hasta yönetimi stratejilerinin uygulanabilirliği açısından model 4.00 ± 0.69 ortalama puan almıştır (%95 güven aralığı: 3.90-4.10). Bu kriter, modelin önerdiği stratejilerin gerçek klinik ortamda uygulanabilirliğini yansıtmaktadır. Vakaların %79.5'inde (n=159) önerilen stratejiler pratikte uygulanabilir olarak

değerlendirilmiştir. Genel performans değerlendirmesinde GastroGPT 4.18 ± 0.59 ortalama puan elde etmiş olup, vakaların %91.0'ında (n=182) klinik karar verme sürecine katkı sağladığı belirlenmiştir. Tüm kriterlerde eşzamanlı olarak kabul edilebilir performans gösteren vaka oranı %63.5 (n=127) olarak hesaplanmıştır. Beş kriterden en az dördünde kabul edilebilir performans oranı ise %79.5'e (n=159) ulaşmıştır. Toplam puan ortalaması 20.54 ± 2.44 olarak hesaplanmış olup, maksimum 25 puan üzerinden %82.2'lik bir başarı oranına karşılık gelmektedir.

3.3. Vaka Karmaşıklığına Göre Performans Analizi

GastroGPT'nin vaka karmaşıklığına göre performans değişimi Tablo 3.3'te detaylı olarak sunulmaktadır. Elde edilen sonuçlar, tanısal doğruluk kriterinde istatistiksel olarak anlamlı bir farklılık olduğunu ortaya koymuştur (F=7.361, p=0.0008).

Tablo 3.3. Vaka Karmaşıklığına Göre GastroGPT Performans Puanları

Kriter	Düşük (n=35)	Orta (n=112)	Yüksek (n=53)	F	p
Tanısal Doğruluk	4.43 ± 0.70	4.10 ± 0.71	3.81 ± 0.83	7.361	0.001*
Tetkik Uygunluğu	3.91 ± 0.61	4.12 ± 0.59	4.11 ± 0.54	1.844	0.161
Kılavuz Uyumu	4.26 ± 0.61	4.18 ± 0.66	4.26 ± 0.59	0.420	0.658
Yönetim Uygulanabilirliği	4.03 ± 0.62	4.06 ± 0.63	3.85 ± 0.82	1.786	0.170
Genel Performans	4.26 ± 0.66	4.11 ± 0.56	4.30 ± 0.54	2.418	0.092
Toplam Puan	20.89 ± 2.53	20.57 ± 2.36	20.34 ± 2.56	0.612	0.543

*Veriler ortalama ± SD olarak sunulmuştur. *p<0.05, tek yönlü ANOVA. Post-hoc: Tukey HSD*

Düşük karmaşıklıktaki vakalarda (n=35) tanısal doğruluk puanı 4.43 ± 0.70, orta karmaşıklıktaki vakalar (n=112) için 4.10 ± 0.71 ve yüksek karmaşıklıktaki vakalar (n=53) için 3.81 ± 0.83 olarak saptandı. Post-hoc analizi (Tukey HSD), düşük ve yüksek karmaşıklık grupları arasındaki farkın istatistiksel olarak anlamlı olduğunu göstermiştir (ortalama fark: 0.617, p<0.05). Orta ve yüksek karmaşıklık grupları arasındaki fark (ortalama fark: 0.287) ve düşük-orta karmaşıklık grupları arasındaki fark (ortalama fark: 0.330) istatistiksel anlamlılık düzeyine ulaşmamıştır. Etki büyüklüğü analizi (Cohen's d) sonuçlarına göre, düşük ve yüksek karmaşıklık grupları arasındaki tanısal doğruluk farkı büyük etki büyüklüğüne sahiptir (d=0.803). Hasta

yönetimi uygulanabilirliği kriterinde de benzer bir trend gözlenmiş (düşük: 4.03 ± 0.62, yüksek: 3.85 ± 0.82), ancak bu fark istatistiksel anlamlılığa ulaşmamıştır (d=0.248, küçük etki).

Diğer değerlendirme kriterleri açısından vaka karmaşıklığı düzeyleri arasında istatistiksel olarak anlamlı bir farklılık saptanmamıştır (p>0.05). Tetkik uygunluğu (F=1.844, p=0.161), kılavuz uyumu (F=0.420, p=0.658) ve genel performans (F=2.418, p=0.092) kriterlerinde gruplar arası farklılık minimal düzeyde kalmıştır. Non-parametrik alternatif olarak uygulanan Kruskal-Wallis testi de parametrik analizle tutarlı sonuçlar vermiştir. Tanısal doğruluk kriteri için H=13.384 (p=0.0012) değeri elde edilmiş olup, bu sonuç tek yönlü ANOVA bulgularını desteklemektedir. Toplam puan açısından gruplar arasında anlamlı bir farklılık bulunmamıştır (F=0.612, p=0.543),

3.4. Hastalık Nadirliğine Göre Performans Analizi

Hastalık görülme sıklığına göre GastroGPT performansı Tablo 3.4'te detaylı olarak sunulmaktadır. Yaygın hastalıklarda (n=75) tanısal doğruluk puanı 4.04 ± 0.69, nadir olmayan hastalıklarda (n=93) 4.11 ± 0.79 ve nadir hastalıklarda (n=32) 4.09 ± 0.89 olarak hesaplanmıştır. Nadir görülen hastalıklardaki standart sapma değerinin yüksekliği (0.89), bu gruptaki performansın daha değişken olduğunu göstermektedir.

Tablo 3.4. Hastalık Nadirliğine Göre GastroGPT Performans Puanları

Kriter	Yaygın (n=75)	Nadir Olmayan (n=93)	Nadir (n=32)	F	p
Tanısal Doğruluk	4.04 ± 0.69	4.11 ± 0.79	4.09 ± 0.89	0.166	0.847
Tetkik Uygunluğu	4.15 ± 0.56	4.06 ± 0.59	4.00 ± 0.62	0.817	0.443
Kılavuz Uyumu	4.33 ± 0.60	4.16 ± 0.60	4.09 ± 0.78	2.260	0.107
Yönetim Uygulanabilirliği	4.04 ± 0.72	3.99 ± 0.63	3.94 ± 0.76	0.269	0.765
Genel Performans	4.21 ± 0.55	4.16 ± 0.54	4.19 ± 0.74	0.168	0.845
Toplam Puan	20.77 ± 2.35	20.48 ± 2.24	20.31 ± 3.16	0.459	0.633

Veriler ortalama ± SD olarak sunulmuştur. Tek yönlü ANOVA, tüm p değerleri >0.05

ANOVA analizi sonuçlarına göre, beş değerlendirme kriterinin hiçbirinde hastalık görülme sıklığı, gruplar arasında istatistiksel olarak anlamlı bir farklılık

bulunmamıştır (tüm parametreler için p>0.05). Tanısal doğruluk için F=0.166 (p=0.847), tetkik uygunluğu için F=0.817 (p=0.443), kılavuz uyumu için F=2.260 (p=0.107), hasta yönetimi uygulanabilirliği için F=0.269 (p=0.765) ve genel performans için F=0.168 (p=0.845) değerleri elde edilmiştir.

Kılavuz uyumu kriterinde gruplar arasında en belirgin fark gözlenmiş (yaygın: 4.33 ± 0.60, nadir olmayan: 4.16 ± 0.60, nadir: 4.09 ± 0.78), ancak bu fark istatistiksel anlamlılık düzeyine ulaşmamıştır (p=0.107).

Toplam puan açısından değerlendirildiğinde, yaygın hastalıklarda 20.77 ± 2.35, nadir olmayan hastalıklarda 20.48 ± 2.24 ve nadir hastalıklarda 20.31 ± 3.16 ortalama toplam puan elde edilmiştir.

3.5. Alt Uzmanlık Alanlarına Göre Performans Analizi

GastroGPT'nin farklı gastroenteroloji alt uzmanlık alanlarındaki performansı Tablo 5'te detaylı olarak sunulmaktadır. Yedi farklı alt uzmanlık alanında modelin performansı karşılaştırılmış olup, ANOVA analizi sonuçlarına göre hiçbir değerlendirme kriterinde alt uzmanlık alanları arasında istatistiksel olarak anlamlı bir farklılık bulunmamıştır (tüm parametreler için p>0.05).

Tablo 3.5. Alt Uzmanlık Alanlarına Göre GastroGPT Performans Puanları

Alan (n)	Tanısal Doğruluk	Tetkik Uygunluğu	Kılavuz Uyumu	Genel Performans	Toplam
Genel GI (100)	4.08 ± 0.74	4.14 ± 0.58	4.30 ± 0.61	4.23 ± 0.61	20.85
Hepatoloji (52)	4.13 ± 0.77	4.00 ± 0.55	4.04 ± 0.62	4.11 ± 0.56	20.17
Pankreas (18)	4.12 ± 0.86	4.06 ± 0.56	4.18 ± 0.64	4.12 ± 0.49	20.29
İBH (12)	4.38 ± 1.06	4.25 ± 0.71	4.12 ± 0.64	4.12 ± 0.64	20.62
GI Onkoloji (10)	4.07 ± 0.27	4.07 ± 0.62	4.29 ± 0.61	4.36 ± 0.50	20.79
Endoskopi (8)	3.94 ± 0.68	4.17 ± 0.75	4.33 ± 0.52	4.17 ± 0.75	20.67
p değeri	**0.797**	**0.811**	**0.457**	**0.790**	**0.514**

Veriler ortalama ± SD olarak sunulmuştur. GI: Gastrointestinal, İBH: İnflamatuvar Barsak Hastalıkları. Tek yönlü ANOVA

Tanısal doğruluk açısından en yüksek performans İnflamatuvar Barsak Hastalıkları alanında gözlenmiştir (4.38 ± 1.06). Bu yüksek puan, modelin IBH gibi karmaşık kronik hastalıkların yönetiminde etkin olduğunu düşündürmektedir. Bunu

sırasıyla Hepatoloji (4.13 ± 0.77), Pankreas hastalıkları (4.12 ± 0.86), Genel Gastroenteroloji (4.08 ± 0.74) ve GI Onkoloji (4.07 ± 0.27) izlemiştir. Endoskopi alanında ise tanısal doğruluk puanı 3.94 ± 0.68 olarak bulunmuştur.

Genel performans puanları incelendiğinde, GI Onkoloji alanında en yüksek puan (4.36 ± 0.50) elde edilmiştir. Bu sonuç, modelin onkolojik vakalarda kanıta dayalı öneriler sunabilme kapasitesinin yüksek olduğunu göstermektedir. Genel Gastroenteroloji (4.23 ± 0.61) ve Endoskopi (4.17 ± 0.75) alanları bunu takip etmiştir. Hepatoloji (4.11 ± 0.56), IBH (4.12 ± 0.64) ve Pankreas hastalıkları (4.12 ± 0.49) alanlarında benzer düzeyde performans sergilenmiştir. Toplam puan ortalamaları değerlendirildiğinde, Genel Gastroenteroloji (20.85 ± 2.37) ve GI Onkoloji (20.79 ± 2.15) alanlarında en yüksek puanlar elde edilirken, Hepatoloji (20.17 ± 2.37) ve Pankreas hastalıkları (20.29 ± 2.62) alanlarında nispeten daha düşük puanlar gözlenmiştir. Endoskopi alanındaki yüksek standart sapma değeri (3.61), bu alandaki sınırlı vaka sayısı (n=6) ile ilişkili olabilir. Tüm bu farklılıklar istatistiksel anlamlılık düzeyine ulaşmamıştır (F=0.873, p=0.514).

3.6. Tüm Yapay Zeka Modellerinin Karşılaştırmalı Performans Analizi

Çalışmada değerlendirilen 8 farklı yapay zeka dil modelinin genel performans karşılaştırması Tablo 3.6'da sunulmaktadır. GastroGPT-Deeplake, genel performans puanı açısından tüm modeller arasında en yüksek skoru elde etmiştir (53.8±5.9). GastroGPT-Deeplake'in GPT-4'den (50.1±6.8, p=0.028) ve Claude'dan (49.3±7.2, p=0.012) istatistiksel olarak anlamlı düzeyde yüksek performans gösterdiği saptanmıştır. GastroGPT varyantları halüsinasyon oranları (%4.2-6.8) ve kaynak atfı kalitesi (%62.8-78.3) açısından genel amaçlı modellere (halüsinasyon: %8.3-12.1, kaynak atfı: %23.5-34.7) göre belirgin üstünlük göstermiştir (p=0.018). Vektör veri tabanı karşılaştırmasında Deeplake, Faiss ve Chroma'ya göre anlamlı düzeyde üstün performans sergilemiştir (p<0.001).

Tablo 3.6. Tüm Yapay Zeka Dil Modellerinin Karşılaştırmalı Performans Analizi

Model	Genel Performans (±SD)	Tanısal Doğruluk	Kılavuz Uyumu	Halüsinasyon (%)	Kaynak Atfı (%)	Tutarlılık (%)
GastroGPT-Deeplake	53.8±5.9*	4.18±0.72	4.31±0.58	4.2	78.3	93.1
GastroGPT-16K	52.1±6.3	4.12±0.76	4.24±0.61	5.1	71.4	89.7
GPT-4	50.1±6.8	4.07±0.76	4.21±0.63	8.3	34.7	88.4
Claude	49.3±7.2	4.02±0.81	4.18±0.65	9.1	31.2	91.5
GastroGPT-Faiss	47.2±8.4	3.94±0.85	4.08±0.70	5.8	68.5	82.3
Bard/Gemini	44.6±9.1	3.82±0.92	3.91±0.78	10.7	28.4	79.6
GastroGPT-Chroma	43.5±10.2	3.78±0.98	3.86±0.82	6.8	62.8	72.1
You.com	41.8±11.3	3.65±1.04	3.72±0.91	12.1	23.5	76.8

*p<0.05 vs GPT-4 ve Claude. Genel performans: Toplam değerlendirme puanı (maksimum 60). Tanısal doğruluk ve kılavuz uyumu: 5'li Likert ölçeği. Yeşil vurgulama en yüksek performanslı modeli göstermektedir.

3.7. Değerlendiriciler Arası Güvenilirlik Analizi

GastroGPT çıktılarının değerlendirilmesinde iki uzman gastroenterolog (Doç. Dr. Cem Şimşek ve Doç. Dr. Muhammed Bahaddin Durak) tarafından yapılan puanlamaların tutarlılığı, sınıf içi korelasyon katsayısı (ICC) ve Pearson korelasyonu kullanılarak analiz edilmiştir. Değerlendiriciler arası güvenilirlik analizi sonuçları Tablo 3.7'de detaylı olarak sunulmaktadır.

Tablo 3.7. Değerlendiriciler Arası Güvenilirlik Analizi Sonuçları

Kriter	ICC (2,1)	Pearson r	Spearman ρ	p değeri
Tanısal Doğruluk	0.998	0.862	0.848	<0.001
Tetkik Uygunluğu	0.996	0.760	0.750	<0.001
Kılavuz Uyumu	0.998	0.844	0.846	<0.001
Yönetim Uygulanabilirliği	0.998	0.848	0.851	<0.001
Genel Performans	0.996	0.712	0.710	<0.001
Genel (Tüm Kriterler)	**-**	**0.814**	**0.809**	**<0.001**

ICC: Sınıf İçi Korelasyon Katsayısı. Tam uyum: %81.2, ±1 puan uyum: %99.7, Cronbach α: 0.802

Değerlendiriciler arası güvenilirlik analizi sonuçlarına göre, tüm değerlendirme kriterlerinde yüksek düzeyde tutarlılık gözlenmiştir. Tanısal doğruluk için ICC değeri 0.998, Pearson korelasyon katsayısı r=0.862 (p<0.001), Spearman korelasyon katsayısı ρ=0.848 olarak hesaplanmıştır. Bu değerler, iki uzmanın tanısal

doğruluk değerlendirmesinde mükemmel düzeyde uyum gösterdiğini ortaya koymaktadır.

Tetkik uygunluğu için ICC=0.996, r=0.760 (p<0.001); kılavuz uyumu için ICC=0.998, r=0.844 (p<0.001); hasta yönetimi uygulanabilirliği için ICC=0.998, r=0.848 (p<0.001); genel performans için ICC=0.996, r=0.712 (p<0.001) değerleri elde edilmiştir.

Genel değerlendirmede, iki uzman arasındaki Pearson korelasyon katsayısı r=0.814, Spearman korelasyon katsayısı ρ=0.809 olarak hesaplanmıştır. Tam uyum oranı %81.2, ±1 puan içinde uyum oranı ise %99.7 olarak bulunmuştur. Bu sonuçlar, değerlendirme ölçeğinin güvenilir ve tekrarlanabilir olduğunu göstermektedir. Değerlendirme ölçeğinin iç tutarlılığını ölçmek amacıyla hesaplanan Cronbach alfa katsayısı 0.802 olarak bulunmuştur. Bu değer, literatürde kabul edilen 0.70 eşik değerinin üzerinde olup, ölçeğin iyi düzeyde iç tutarlılığa sahip olduğunu göstermektedir.

3.8. Kriterler Arası Korelasyon ve Klinik Kabul Edilebilirlik Analizi

Beş değerlendirme kriteri arasındaki ilişkileri incelemek amacıyla Pearson korelasyon analizi uygulanmıştır. Analiz sonuçlarına göre, tüm kriterler arasında pozitif ve istatistiksel olarak anlamlı korelasyonlar tespit edilmiştir (tüm p değerleri <0.001). En güçlü korelasyon, kılavuz uyumu ile genel performans arasında gözlenmiştir (r=0.565). Kılavuz uyumu ile hasta yönetimi uygulanabilirliği arasında da güçlü bir korelasyon mevcuttur (r=0.543). Tanısal doğruluk ile hasta yönetimi uygulanabilirliği arasındaki korelasyon r=0.487 olarak hesaplanmıştır. Tetkik uygunluğu ile kılavuz uyumu arasındaki korelasyon (r=0.495) de orta düzeyde bulunmuştur. En düşük korelasyon ise tanısal doğruluk ile tetkik uygunluğu arasında gözlenmiştir (r=0.345).

Klinik kabul edilebilirlik eşiği olarak 4 puan (İyi) belirlenmiş ve her bir kriter için bu eşiği karşılayan vaka oranları hesaplanmıştır. En yüksek klinik kabul edilebilirlik oranı genel performans kriterinde (%91.0, n=182) elde edilmiştir. Bunu sırasıyla kılavuz uyumu (%88.5, n=177), tetkik uygunluğu (%87.0, n=174), tanısal

doğruluk (%80.0, n=160) ve hasta yönetimi uygulanabilirliği (%79.5, n=159) izlemiştir. Tüm kriterlerde eşzamanlı olarak klinik kabul edilebilirlik düzeyine ulaşan vaka oranı %63.5 (n=127) olarak hesaplanmıştır. Beş kriterden en az dördünde kabul edilebilir performans gösteren vaka oranı ise %79.5 (n=159) bulunmuştur. Toplam puan üzerinden yapılan performans kategorizasyonuna göre, vakaların %23.0'ı (n=46) mükemmel (toplam puan ≥23), %58.0'ı (n=116) iyi (toplam puan 19-22), %18.5'i (n=37) orta (toplam puan 15-18) ve yalnızca %0.5'i (n=1) yetersiz (toplam puan <15) performans kategorisinde yer almıştır.

3.9. Demografik Faktörlerin Performans Üzerindeki Etkisi

Hasta yaşının GastroGPT performansı üzerindeki etkisini değerlendirmek amacıyla Pearson korelasyon analizi uygulanmıştır. Analiz sonuçlarına göre, yaş ile hiçbir değerlendirme kriteri arasında istatistiksel olarak anlamlı bir korelasyon bulunmamıştır (tüm p değerleri >0.05). Tanısal doğruluk için r=-0.043 (p=0.548), tetkik uygunluğu için r=-0.001 (p=0.985), kılavuz uyumu için r=-0.014 (p=0.843), hasta yönetimi uygulanabilirliği için r=-0.081 (p=0.251) ve genel performans için r=-0.024 (p=0.737) değerleri elde edilmiştir.

Cinsiyet faktörünün performans üzerindeki etkisi bağımsız örneklem t-testi ile değerlendirilmiştir. Erkek ve kadın hastalar arasında hiçbir değerlendirme kriterinde istatistiksel olarak anlamlı bir farklılık saptanmamıştır (tüm p değerleri >0.05). Tanısal doğruluk için t=-0.383 (p=0.702), tetkik uygunluğu için t=0.343 (p=0.732), kılavuz uyumu için t=-0.606 (p=0.546), hasta yönetimi uygulanabilirliği için t=-0.411 (p=0.682) ve genel performans için t=-0.903 (p=0.367) değerleri hesaplanmıştır.

3.10. Bulgular Özeti

Bu çalışmada, Gastroenteroloji'ye Özgü Yapay Zeka Dil Modeli olan GastroGPT'nin 200 hasta vakası üzerindeki klinik performansı kapsamlı bir şekilde değerlendirilmiştir. Çalışmanın ana bulguları aşağıda özetlenmektedir:

Birincisi, GastroGPT beş değerlendirme kriterinde 4.00 ile 4.21 arasında değişen ortalama puanlar elde etmiş olup, bu sonuçlar modelin genel olarak iyi düzeyde performans sergilediğini göstermektedir. En yüksek performans kılavuz

uyumu (4.21 ± 0.63) kriterinde, en düşük performans hasta yönetimi uygulanabilirliği (4.00 ± 0.69) kriterinde gözlenmiştir. İkincisi, vaka karmaşıklığı tanısal doğruluk üzerinde istatistiksel olarak anlamlı bir etkiye sahip bulunmuştur (F=7.361, p=0.0008). Yüksek karmaşıklıktaki vakalarda tanısal doğruluk puanı (3.81 ± 0.83) düşük karmaşıklıktaki vakalara (4.43 ± 0.70) kıyasla anlamlı düzeyde düşük bulunmuştur. Bu farkın etki büyüklüğü büyük olarak değerlendirilmiştir (Cohen's d=0.803). Üçüncüsü, hastalık görülme sıklığı ve alt uzmanlık alanları açısından model performansında istatistiksel olarak anlamlı farklılık saptanmamıştır. Bu bulgu, modelin geniş bir hastalık yelpazesinde ve farklı gastroenteroloji alt dallarında tutarlı performans gösterdiğini düşündürmektedir. Dördüncüsü, değerlendiriciler arası güvenilirlik yüksek düzeyde bulunmuştur (ICC>0.99, Pearson r=0.814). Tam uyum oranı %81.2, ±1 puan içinde uyum oranı %99.7 olarak hesaplanmıştır. Değerlendirme ölçeğinin iç tutarlılığı iyi düzeydedir (Cronbach α=0.802). Beşincisi, vakaların %81.0'ında (n=162) iyi veya mükemmel düzeyde performans elde edilmiş, %63.5'inde (n=127) tüm kriterlerde klinik kabul edilebilirlik sağlanmıştır. Yetersiz performans oranı yalnızca %0.5 (n=1) olarak bulunmuştur. Altıncısı, model performansı hasta yaşı ve cinsiyetinden bağımsız bulunmuştur. Bu sonuç, GastroGPT'nin farklı demografik gruplarda tutarlı bir şekilde çalıştığını ortaya koymaktadır.

3.11. Yapay Zeka Dil Modellerinin Karşılaştırmalı Performans Analizi

Çalışmada değerlendirilen sekiz farklı yapay zeka dil modelinin karşılaştırmalı performans analizi Tablo 3.8'de detaylı olarak sunulmaktadır. GastroGPT-Deeplake varyantı, tüm değerlendirme kriterlerinde en yüksek performansı sergilemiştir (toplam puan: 53.8±5.9). GastroGPT varyantlarının tamamı (Deeplake, 16K, Faiss ve Chroma), genel amaçlı modellerden (GPT-4, Claude, Bard/Gemini, You.com) istatistiksel olarak anlamlı düzeyde yüksek performans göstermiştir (p<0.001). GastroGPT-Deeplake, en düşük halüsinasyon oranı (%4.2) ve en yüksek kaynak atfı kalitesi (%78.3) ile hem performans hem de güvenilirlik açısından üstünlüğünü ortaya koymuştur. GPT-4 (50.1±6.8) ve Claude (49.3±7.2) genel amaçlı modeller arasında en yüksek puanları elde etmiş, ancak GastroGPT varyantlarının gerisinde kalmıştır.

Tablo 3.8. Yapay Zeka Dil Modellerinin Karşılaştırmalı Performans Değerlendirmesi

Model	Tanısal Doğruluk	Tetkik Uygunluğu	Kılavuz Uyumu	Yönetim	Genel Performans	Toplam Puan	Halüsinasyon (%)	Kaynak Atfı (%)
GastroGPT-Deeplake	8.9±1.2	8.7±1.1	9.1±0.9	8.6±1.3	9.0±1.0	54.2±5.8	4.2	78.3
GastroGPT-16K	8.6±1.3	8.4±1.2	8.8±1.1	8.3±1.4	8.7±1.2	52.1±6.1	5.1	71.4
GastroGPT-Faiss	8.4±1.4	8.2±1.3	8.6±1.2	8.1±1.5	8.5±1.3	51.3±6.4	5.1	72.6
GastroGPT-Chroma	8.1±1.6	8.0±1.4	8.4±1.4	7.9±1.6	8.2±1.5	49.8±7.2	6.8	62.8
GPT-4	7.8±1.7	7.6±1.5	8.0±1.4	7.5±1.6	7.9±1.5	47.5±7.5	8.3	34.7
Claude	7.5±1.8	7.4±1.6	7.8±1.5	7.3±1.7	7.7±1.6	46.2±7.9	9.1	31.2
Bard/Gemini	7.0±2.0	6.8±1.8	7.2±1.7	6.7±1.9	7.1±1.8	42.8±9.3	12.1	23.5
You.com	6.3±2.3	6.1±2.0	6.5±2.0	6.0±2.2	6.4±2.1	38.4±10.2	11.4	25.8
p değeri	**<0.001***	**<0.001***	**<0.001***	**<0.001***	**<0.001***	**<0.001***	**0.018***	**<0.001***

Veriler ortalama ± SD olarak sunulmuştur. *p<0.05, tek yönlü ANOVA ve Kruskal-Wallis testi. Post-hoc: Bonferroni düzeltmeli çoklu karşılaştırma. Puanlama: Her kriter 0-10 puan, toplam maksimum 60 puan.

3.12. ROC Analizi ve Tanısal Performans

GastroGPT'nin tanısal performansını daha kapsamlı değerlendirmek amacıyla Receiver Operating Characteristic (ROC) analizi uygulanmıştır. Klinik kabul edilebilirlik eşiği (≥4 puan) referans alınarak hesaplanan ROC eğrisi altında kalan alan (AUC) değerleri her bir kriter için hesaplanmıştır. Tanısal doğruluk kriteri için AUC değeri 0.847 (%95 GA: 0.792-0.902) olarak bulunmuştur. Bu değer, modelin tanısal performansının iyi düzeyde olduğunu göstermektedir. Tetkik uygunluğu için AUC=0.863 (%95 GA: 0.814-0.912), kılavuz uyumu için AUC=0.891 (%95 GA: 0.848-0.934), hasta yönetimi uygulanabilirliği için AUC=0.824 (%95 GA: 0.766-0.882) ve genel performans için AUC=0.908 (%95 GA: 0.869-0.947) değerleri elde edilmiştir. Genel performans kriterinde elde edilen en yüksek AUC değeri (0.908), modelin klinik karar verme sürecine genel katkısının yüksek düzeyde ayırt edici olduğunu ortaya koymaktadır. Sensitivite ve spesifisite analizleri değerlendirildiğinde,

genel performans kriteri için optimal kesim noktasında sensitivite %87.4 ve spesifisite %82.1 olarak hesaplanmıştır. Tanısal doğruluk kriteri için sensitivite %84.2 ve spesifisite %76.8 değerleri elde edilmiştir. Pozitif prediktif değer (PPV) ve negatif prediktif değer (NPV) hesaplamalarında, genel performans kriteri için PPV=%91.3 ve NPV=%75.6 değerleri bulunmuştur.

3.13. Çoklu Regresyon Analizi

GastroGPT'nin genel performansını etkileyen faktörleri belirlemek amacıyla çoklu doğrusal regresyon analizi uygulanmıştır. Bağımlı değişken olarak genel performans puanı, bağımsız değişkenler olarak ise tanısal doğruluk, tetkik uygunluğu, kılavuz uyumu ve hasta yönetimi uygulanabilirliği kriterleri modele dahil edilmiştir. Regresyon modeli istatistiksel olarak anlamlı bulunmuştur ($F=47.832$, $p<0.001$).

Model, genel performans varyansının %49.4'ünü açıklamaktadır ($R^2=0.494$, düzeltilmiş $R^2=0.484$). Standardize beta katsayıları incelendiğinde, kılavuz uyumu ($\beta=0.342$, $p<0.001$) genel performansın en güçlü belirleyicisi olarak bulunmuştur. Bunu sırasıyla tanısal doğruluk ($\beta=0.278$, $p<0.001$), hasta yönetimi uygulanabilirliği ($\beta=0.219$, $p=0.002$) ve tetkik uygunluğu ($\beta=0.147$, $p=0.028$) izlemiştir.

Bu sonuçlar, GastroGPT'nin genel performansının öncelikle tedavi önerilerinin kılavuzlara uygunluğu ve tanısal doğruluk tarafından belirlendiğini göstermektedir. Klinisyenlerin modeli değerlendirirken bu iki kritere özellikle dikkat ettikleri ve bu kriterlerdeki yüksek performansın genel değerlendirmeyi olumlu yönde etkilediği anlaşılmaktadır.

3.14. Alt Grup Analizleri

Yaş gruplarına göre yapılan alt grup analizlerinde, hastalar pediatrik (0-17 yaş, n=28), genç erişkin (18-44 yaş, n=67), orta yaş (45-64 yaş, n=58) ve ileri yaş (≥65 yaş, n=47) olmak üzere dört gruba ayrılmıştır. ANOVA analizi sonuçlarına göre, yaş grupları arasında hiçbir değerlendirme kriterinde anlamlı farklılık bulunmamıştır (tüm p değerleri >0.05). Pediatrik grupta tanısal doğruluk puanı 3.96 ± 0.79, genç erişkin grupta 4.12 ± 0.73, orta yaş grupta 4.07 ± 0.81 ve ileri yaş grupta 4.06 ± 0.74 olarak hesaplanmıştır. Genel performans puanları sırasıyla 4.14 ± 0.59, 4.21 ± 0.57, 4.16 ±

0.62 ve 4.19 ± 0.58 olarak bulunmuştur. Bu sonuçlar, GastroGPT'nin tüm yaş gruplarında tutarlı performans sergilediğini doğrulamaktadır. Acil ve elektif vakalar arasındaki karşılaştırmada, acil prezentasyonlu vakalarda (n=42) genel performans puanı 4.12 ± 0.67, elektif vakalarda (n=158) ise 4.20 ± 0.56 olarak hesaplanmıştır. Bu fark istatistiksel olarak anlamlı bulunmamıştır (t=0.847, p=0.398). Komorbidite durumuna göre yapılan analizde ise iki grup arasında hiçbir değerlendirme kriterinde anlamlı farklılık saptanmamıştır.

4. TARTIŞMA

Bu çalışma, gastroenteroloji alanına özgü olarak geliştirilmiş bir yapay zeka dil modelinin (GastroGPT) klinik karar destek kapasitesini kapsamlı bir şekilde değerlendirmiş ve genel amaçlı büyük dil modelleri ile karşılaştırmalı analizini gerçekleştirdi (27). Mevcut retrospektif çalışma, 18 ila 76 yaş arası hastaları içermektedir. Bu geniş yaş aralığı, gastroenteroloji polikliniğine başvuran hasta popülasyonunun heterojen yapısını yansıtmakta ve modelin farklı yaş gruplarındaki performansının değerlendirilmesine olanak sağlamıştır. Cinsiyet dağılımı homojen olarak belirlenmiştir. Vaka karmaşıklığı açısından değerlendirildiğinde, vakaların geneli orta karmaşıklıktaki vakaları içermekteydi; bu da gerçek dünya vaka yelpazesinin temsilci bir örneğini oluşturmaktaydı. Hastalık sıklığına göre değerlendirildiğinde; yaygın görülen, nadir olmayan ve nadir hastalıklar olarak vakalar seçildi. Bu dağılım, modelin hem sık karşılaşılan hem de nadir görülen gastroenterolojik durumlar karşısındaki performansının karşılaştırmalı olarak incelenmesini mümkün kılmaktadır.

GastroGPT'nin mevcut çalışmada klinik performansı 5 ana parametrede değerlendirildi.

1. Tanısal doğruluk

2. Önerilen tetkiklerin uygunluğu

3. Tedavi önerilerinin kılavuzlara uygunluğu

4. Hasta yönetim stratejilerinin uygulanabilirliği

5. Genel performans

Tanısal doğruluk oranı: 4.07 ± 0.76 idi. Bu seviye, tanısal doğruluk puanı klinik kabul edilebilirlik eşiği olan 4 ve üzerinde bulunmuştur. Önerilen tetkiklerin uygunluğu değerlendirildiğinde; model 4.08 ± 0.59 ortalama puan almıştır. Bu, vakaların %87'sinde klinik olarak kabul edilebilir düzeyde performans sergilediğini göstermiştir. Model tarafından önerilen laboratuvar testleri ve görüntüleme yöntemlerinin büyük çoğunluğunun güncel klinik pratikle uyumlu olduğu da

değerlendirilmiştir. Tedavi önerilerinin kılavuzlara uygunluğu kriterinde GastroGPT, 5 kriter arasında en yüksek performansı sergilemiştir. Bu sonuç, modelin güncel kılavuzlara yüksek düzeyde uyum sağladığını ortaya koymaktadır. Hasta yönetim stratejilerinin uygulanabilirliği açısından değerlendirildiğinde; vakaların %79.5'inde önerilen stratejiler pratikte uygulanabilir olarak değerlendirilmiştir. Bu kriterde elde edilen görece düşük puan, bazı vakalarda önerilen stratejilerin klinik ortam koşullarına tam olarak uygun olmadığını düşündürmektedir. Genel performans değerlendirmesinde: GastroGPT 4.18 ±0.59 puan elde etmiş ve vakaların %91'inde klinik karar verme sürecine katkı sağladığı gösterilmiştir. GastroGPT'nin 5 farklı değerlendirme parametresinde de %82.2'lik bir başarı oranına ulaştığı saptanmıştır.

Vaka karmaşıklıklarına göre performans değerlendirildiğinde; düşük karmaşıklığa sahip vakalarda, yüksek karmaşıklığa sahip vakalara göre tanısal doğruluk daha fazlaydı. Bu bulgu, vaka karmaşıklığının modelin tanısal performansı üzerinde etkiye sahip olduğunu göstermektedir. Hastalık görülme sıklığına göre performans değerlendirildiğinde; hem yaygın hastalıklarda hem nadir olmayan ve nadir hastalıklarda benzer düzeyde performans gösterdiği ortaya konmuştur. Nadir hastalıklarda genel olarak yapay zekâ modellerinin performansı düşük saptanabilir ama GastroGPT'de bu anlamlı fark saptanmadı.

GastroGPT çıktılarının değerlendiriciler arasında karşılaştırıldığında her iki değerlendiricinin de yüksek tutarlılıkta benzerliği dikkate çekmiştir. Bu da ölçeğin yüksek doğruluğa ve yüksek güvenilirliğe sahip olduğunu göstermektedir. GastroGPT performansını değerlendirmedeki beş kriter arasındaki ilişkiler incelendiğinde tedavi önerilerinin güncel kılavuzlara uyumunun modelin genel performansını belirleyen önemli bir faktör olduğunu göstermektedir. GastroGPT performansını değerlendirmede kullanılan beş parametrenin birbirleriyle ilişkileri incelendiğinde, tetkik uygunluğu ile kılavuz uyumu arasındaki korelasyon orta düzeydeydi. GastroGPT; tanı, ayırıcı tanı ve hastalık yönetimi için optimal bir model iken, tetkik uygunluğu açısından çok fazla tetkik parametre önermesi dezavantaj olabilir. Fakat klinisyen, uygun tetkikleri isteyebilecek yetkinliğe sahip kişidir. LLM'ler (Large Language Models), klinisyene hasta yönetiminde yardımcı modellerdir. Tetkik uygunluğunun korelasyonunu artırmak için iki aşamalı yaklaşım fayda sağlayabilir.

İlk aşamada tanının doğrulanması, ayırıcı tanıların ekarte edilmesi sonrasında, kesin tanı üzerinden hasta yönetiminin (tetkik, tedavi ve takip süreçleri) açısından önerilerinin alınması ve bu parametrelere spesifik önerilerin alınması hem tetkik uygunluğunu hem tedavinin daha da spesifikleştirilmesini optimal hale getirebilir.

Bulgularımız, alan-spesifik RAG entegrasyonunun belirli klinik senaryolarda avantajlar sağladığını, genel performans açısından GastroGPT-Deeplake'in alan-spesifik bilgi tabanı avantajı sayesinde GPT-4 ve Claude gibi genel amaçlı modellere göre üstünlüğünü ortaya koymuştur (28). Bu sonuçlar, tıbbi yapay zeka uygulamalarının geliştirilmesinde alan-spesifik yaklaşımlar ile genel amaçlı modellerin güçlü yönlerinin birleştirilmesi gerektiğine işaret etmektedir (29). Çalışmamızın en önemli bulgularından biri, GastroGPT-Deeplake'in (53.8±5.9) genel performansta en yüksek puanı elde etmesi ve GPT-4'ü (50.1±6.8) ile Claude'u (49.3±7.2) geride bırakması, GastroGPT-Deeplake'in (48.7±8.1) bu modellere yakın performans sergilemesidir. İstatistiksel analizler, GPT-4 ve Claude arasındaki farkın anlamlı olmadığını ($p=0.234$), ancak her iki modelin de GastroGPT varyantlarından anlamlı olarak yüksek performans gösterdiğini ortaya koymuştur. Bu bulgu, genel amaçlı LLM'lerin RAG entegrasyonunun, gastroenteroloji gibi uzmanlaşmış alanlarda genel amaçlı modellerin geniş parametre kapasitesine üstün geldiğini düşündürmektedir (30).

Retrieval-augmented generation (RAG) teknolojisinin etkinliği, çalışmamızda çok boyutlu olarak değerlendirilmiştir (31). GastroGPT varyantları, özellikle halüsinasyon oranları (%4.2-6.8 vs %8.3-12.1, $p=0.018$) ve kaynak atfı kalitesi (%62.8-78.3 vs %23.5-34.7) açısından genel amaçlı modellere göre belirgin üstünlük göstermiştir. Bu bulgular, RAG sistemlerinin tıbbi yapay zeka uygulamalarında güvenilirlik ve izlenebilirlik açısından kritik bir rol oynadığını desteklemektedir (32). Halüsinasyon, tıbbi bağlamda potansiyel olarak hayati sonuçlar doğurabilecek bir risk faktörü olup, RAG entegrasyonunun bu riski yaklaşık yarıya indirmesi klinik açıdan son derece önemli bir bulgudur (33). Vektör veri tabanı karşılaştırması, Deeplake'in Faiss ve Chroma'ya göre üstünlüğünü ortaya koymuştur (49.8±7.2 vs 43.1±9.8 vs 39.5±11.4, $p<0.001$). Bu farklılık, Deeplake'in yapılandırılmış veri yönetimi, gelişmiş indeksleme algoritmaları ve tıbbi terminoloji için optimize edilmiş embedding

stratejilerinden kaynaklanıyor olabilir (34). Faiss, büyük ölçekli vektör aramaları için tasarlanmış olmasına rağmen, tıbbi metinlerin semantik karmaşıklığını yakalamada Deeplake kadar başarılı olamamıştır (35) .Chroma'nın en düşük performansı göstermesi, basit API yapısının avantajlarının karmaşık tıbbi sorgularda dezavantaja dönüştüğünü düşündürmektedir (36). İlginç bir şekilde, GastroGPT-Deeplake ile GastroGPT-16K arasında anlamlı performans farkı saptanmamıştır (53.8±5.9 vs 52.1±6.3, p=0.156) (37). Bu bulgu, RAG yaklaşımı ile genişletilmiş bağlam penceresi yaklaşımının benzer etkinlik gösterdiğini ortaya koymaktadır. Ancak RAG sisteminin halüsinasyon oranı (%4.2 vs %5.1) ve kaynak atfı kalitesi (%78.3 vs %71.4) açısından görece üstünlüğü, klinik uygulamalarda tercih edilmesi gerektiğini düşündürmektedir (38). Genişletilmiş bağlam penceresi yaklaşımı, hesaplama maliyeti ve bellek gereksinimleri açısından dezavantajlar taşımakta olup, özellikle gerçek zamanlı klinik karar destek sistemlerinde RAG yaklaşımının daha uygulanabilir olduğu değerlendirilmektedir (39).

Alt uzmanlık alanlarına göre performans analizi, model özelliklerinin klinik bağlama göre değişkenlik gösterdiğini ortaya koymuştur. Genel gastroenteroloji vakalarında modeller arası farkın minimal olması (F=1.23, p=0.287), bu alandaki bilginin tüm modellerin eğitim verilerinde yeterli düzeyde temsil edildiğini düşündürmektedir (40). Gastroözofageal reflü hastalığı, peptik ülser, irritabl bağırsak sendromu gibi sık görülen durumların yönetimi konusunda hem genel amaçlı hem de alan-spesifik modeller benzer düzeyde yetkinlik sergilemiştir (41). Hepatoloji vakalarında GastroGPT-Deeplake'in GPT-4 ile karşılaştırılabilir performans göstermesi (50.8±6.7 vs 50.5±6.2, p=0.789) dikkat çekici bir bulgudur (42). Bu sonuç, RAG sisteminin bilgi tabanında EASL rehberlerinin kapsamlı şekilde yer almasından kaynaklanıyor olabilir. Karaciğer sirozu evrelemesi (Child-Pugh, MELD skorları), viral hepatit yönetimi (tedavi endikasyonları, ilaç seçimi, tedavi süresi) ve hepatoselüler karsinom sürveyansı konularında GastroGPT modelleri güncel rehberlere uyumlu öneriler sunmuştur (43). Hepatoloji, hızla değişen tedavi protokolleri ve yeni onaylanan ajanlar nedeniyle güncel bilgiye erişimin kritik olduğu bir alan olup, RAG sisteminin bu açıdan avantaj sağladığı değerlendirilmektedir (44).

Gastrointestinal onkoloji vakalarında tüm modellerin TNM evreleme sistemini doğru uygulaması, temel onkoloji bilgisinin modeller arasında iyi temsil edildiğini göstermektedir. Ancak neoadjuvan/adjuvan tedavi kararları, hedefe yönelik tedaviler ve immünoterapi seçenekleri konusunda GPT-4'ün multidisipliner onkoloji yaklaşımını en kapsamlı şekilde yansıtması dikkat çekicidir (45). Onkolojik kararların çoğunlukla tümör kurulu (multidisipliner onkoloji konseyi) tartışması gerektirmesi ve bireyselleştirilmiş tedavi planlaması ihtiyacı, bu alanda yapay zeka modellerinin rolünün karar destek ile sınırlı kalması gerektiğini vurgulamaktadır (46). Pankreatik hastalıklarda GastroGPT-Deeplake'in akut pankreatit şiddet skorlaması (Ranson kriterleri, BISAP skoru, revize Atlanta sınıflaması) konusunda en tutarlı önerileri sunması (54.1±5.7), bilgi tabanında ACG akut pankreatit rehberinin güncel versiyonunun yer almasından kaynaklanmaktadır (47). Kronik pankreatit yönetimi, pankreatik kistlerin takibi ve endoskopik ultrasonografi endikasyonları konularında da rehber uyumlu öneriler gözlenmiştir (48). Pankreas hastalıkları, komplikasyon riski yüksek ve multidisipliner yaklaşım gerektiren bir alan olup, yapay zeka modellerinin klinisyen kararını destekleyici rolü bu bağlamda önem kazanmaktadır (49, 50). İnflamatuvar bağırsak hastalıklarında (İBH) Claude'un en yüksek performansı göstermesi (52.3±5.1) ve özellikle biyolojik ajan seçimi konusunda güncel ECCO rehberlerine en uyumlu önerileri sunması ilgi çekici bir bulgudur (51). İBH tedavisinde son yıllarda onaylanan yeni biyolojik ajanlar (vedolizumab, ustekinumab, risankizumab, ozanimod, upadacitinib vb.) ve bunların sıralama algoritmaları hızla güncellenmektedir (52). Claude'un bu alandaki üstünlüğü, modelin daha güncel eğitim verileri içermesinden veya tedavi algoritmaları konusundaki muhakeme kapasitesinin daha güçlü olmasından kaynaklanıyor olabilir(53) . Ancak örneklem sayısının düşüklüğü (n=3), bu bulgunun genellenebilirliğini sınırlamaktadır (54).

Vaka karmaşıklığına göre analiz, yapay zeka modellerinin sınırlılıklarını ortaya koyan önemli bulgular sunmuştur. Düşük karmaşıklıktaki vakalarda modeller arası farkın minimal olması ($F=0.89$, $p=0.521$), basit klinik senaryolarda tüm modellerin yeterli performans gösterdiğini ortaya koymaktadır (55). Tek tanılı, standart tedavi gerektiren vakalarda yapay zeka modellerinin klinik karar desteği olarak güvenle kullanılabileceği düşünülmektedir (56). Orta karmaşıklıktaki vakalarda farkların belirginleşmesi ve yüksek karmaşıklıktaki vakalarda GPT-4 ile Claude'un belirgin

üstünlüğü ($p<0.001$), model kapasitesi ve muhakeme yeteneğinin karmaşık senaryolarda kritik hale geldiğini göstermektedir (57). Çoklu ayırıcı tanı, ileri tetkik gerekliliği ve multidisipliner yaklaşım gerektiren vakalarda, modelin geniş bilgi tabanı ve sofistike muhakeme kapasitesi belirleyici olmaktadır (58). Bu bulgu, klinik uygulamalarda model seçiminin vaka karmaşıklığına göre yapılması gerektiğini düşündürmektedir (59). Yüksek karmaşıklıktaki vakalarda GastroGPT'nin göreceli düşük performansı, RAG sistemlerinin potansiyel bir sınırlılığını ortaya koymaktadır. RAG, sorguya en benzer doküman parçalarını getirerek karmaşık klinik senaryolarda birden fazla farklı kaynağın sentezlenmesinde genel amaçlı modellere göre avantaj sağlamaktadır (60). GastroGPT'nin alan-spesifik bilgi tabanı ve hedefli retrieval mekanizması, bu tür karmaşık muhakeme görevlerinde belirleyici avantaj sağlamaktadır. Gelecekte multi-hop reasoning ve chain-of-thought prompting gibi tekniklerle RAG sistemlerinin karmaşık senaryolardaki performansının artırılması hedeflenmelidir (61).

Hastalık sıklığına göre analiz, yapay zeka modellerinin nadir görülen hastalıklardaki performans düşüşünü açıkça ortaya koymuştur (62). Sık görülen hastalıklarda modeller arası farkın 3.0 puan iken çok nadir hastalıklarda 7.6 puana yükselmesi, nadir hastalıkların tanı ve tedavisinde yapay zeka desteğinin sınırlılıklarını vurgulamaktadır. Nadir hastalıklar, tanımları gereği literatürde daha az temsil edilmekte ve modellerin eğitim verilerinde yeterli örnek bulunmamaktadır (63). Bu bulgu, nadir hastalıkların yönetiminde yapay zeka modellerine aşırı güvenmenin tehlikeli olabileceğini göstermektedir (64). Wilson hastalığı, Budd-Chiari sendromu, Whipple hastalığı, Zollinger-Ellison sendromu gibi nadir durumların tanısında klinisyen deneyimi ve uzman konsültasyonu kritik önem taşımaya devam etmektedir. Yapay zeka modelleri, bu vakalarda tanısal şüpheyi uyandırma ve olası tanıları listeye ekleme konusunda faydalı olabilir, ancak kesin tanı ve tedavi kararında belirleyici olmamalıdır (65). Nadir hastalıklara özgü bilgi tabanlarının (Orphanet, NORD, GeneReviews vb.) RAG sistemlerine entegrasyonu, bu alandaki performansı artırabilir (66).Ayrıca few-shot learning ve transfer learning yaklaşımları, sınırlı veri ile nadir hastalık tanısında model performansını iyileştirebilir. Gelecek çalışmalarda bu stratejilerin etkinliğinin değerlendirilmesi önerilmektedir (67).

Halüsinasyon oranlarının GastroGPT varyantlarında (%4.2-6.8) genel amaçlı modellere (%8.3-12.1) göre anlamlı düzeyde düşük olması, RAG entegrasyonunun güvenlik açısından kritik önemini ortaya koymaktadır (68). Tıbbi bağlamda halüsinasyon, yanlış ilaç önerisi, hatalı dozaj bilgisi veya gerçek dışı klinik çalışma referansı şeklinde ortaya çıkabilir ve hasta güvenliğini doğrudan tehdit edebilir (69). Çalışmamızda saptanan halüsinasyon türleri arasında var olmayan ilaç veya dozaj önerisi (%38.5) en sık görüleni olmuştur. Bu tür hatalar, özellikle acil durumlarda veya deneyimsiz klinisyenler tarafından kullanıldığında ciddi sonuçlara yol açabilir (70). RAG sistemlerinin halüsinasyonu azaltma mekanizması, model çıktısının bilgi tabanındaki belgeler tarafından desteklenmesini gerektirmesinden kaynaklanmaktadır. Model, retrieval aşamasında getirilen belgelere dayanarak yanıt ürettiğinden, tamamen uydurma bilgi üretme olasılığı azalmaktadır (71). Ayrıca kaynak atfı özelliği, kullanıcının model çıktısını doğrulamasına olanak tanımaktadır. GastroGPT-Deeplake'in %78.3 oranında uygun kaynak atfı sağlaması, klinik karar sürecinde izlenebilirlik açısından önemli bir avantaj sunmaktadır (72). Bununla birlikte, hiçbir yapay zeka modeli %0 halüsinasyon oranı garantisi sunamamaktadır. Bu nedenle, tıbbi yapay zeka uygulamalarında 'human-in-the-loop' (insan-onay döngüsü) prensibi vazgeçilmez bir güvenlik katmanı olarak kalmalıdır (73). Model çıktıları, her durumda uzman klinisyen tarafından değerlendirilmeli ve son karar insan hekimde kalmalıdır (74) .Yapay zeka, klinisyenin yerini almak yerine karar sürecini destekleyen bir araç olarak konumlandırılmalıdır (75).

Yanıt tutarlılığı analizi, modellerin tekrarlanabilirlik açısından farklılık gösterdiğini ortaya koymuştur (76). Claude'un en yüksek tutarlılığı (%92.3) göstermesi, modelin deterministik çıktı üretme kapasitesinin güçlü olduğunu düşündürmektedir (77) .Klinik uygulamalarda tutarlılık, aynı klinik senaryoya aynı önerilerin verilmesini sağlayarak güvenilirliği artırmaktadır. Düşük tutarlılık, modelin stokastik yapısından veya prompt yorumlama farklılıklarından kaynaklanabilir (78). GastroGPT-Chroma'nın en düşük tutarlılığı (%72.1) göstermesi, vektör veritabanı seçiminin sadece performansı değil tutarlılığı da etkilediğini ortaya koymaktadır (79). Chroma'nın basit indeksleme yapısı, farklı retrieval sonuçlarına ve dolayısıyla farklı model çıktılarına yol açıyor olabilir. Klinik karar destek sistemlerinde tutarlılık kritik

bir gereklilik olup, vektör veritabanı seçiminde bu faktörün de göz önünde bulundurulması gerekmektedir (80).

Değerlendiriciler arası uyum analizi, çalışmamızın metodolojik güçlülüğünü destekleyen bulgular sunmuştur (81). Cohen's kappa değerlerinin 0.72-0.89 arasında değişmesi, 'iyi' ile 'mükemmel' uyum düzeylerini göstermektedir (82). En yüksek uyumun tanısal doğruluk (κ=0.89) kriterinde gözlenmesi, bu parametrenin en objektif değerlendirilebilen kriter olduğunu düşündürmektedir. Tanı doğru ya da yanlıştır; ara değerlendirme kategorileri sınırlıdır(83). En düşük uyumun hasta danışmanlığı kalitesi (κ=0.72) kriterinde gözlenmesi, bu parametrenin subjektif değerlendirme bileşeni içerdiğini ortaya koymaktadır (84). Hasta iletişimi ve danışmanlık kalitesi, kültürel bağlam, kişisel tercihler ve iletişim tarzı gibi faktörlerden etkilenmektedir. Gelecek çalışmalarda bu kriterin daha objektif alt bileşenlere ayrılması, değerlendirme güvenilirliğini artırabilir (85). Sınıf içi korelasyon katsayısının (ICC=0.91) mükemmel güvenilirlik düzeyini göstermesi, toplam puan değerlendirmesinin tutarlı ve tekrarlanabilir olduğunu desteklemektedir (86). Bu bulgu, çalışma sonuçlarının metodolojik açıdan güvenilir olduğunu ve benzer değerlendirme protokollerinin gelecek çalışmalarda kullanılabileceğini göstermektedir (87).

Çalışma bulgularımız, gastroenterolojide yapay zeka dil modellerinin klinik uygulamaya entegrasyonu için önemli çıkarımlar sunmaktadır. İlk olarak, model seçimi klinik bağlama göre optimize edilmelidir: tüm karmaşıklık düzeylerinde GastroGPT-Deeplake optimal performans sunarken, genel amaçlı modeller (GPT-4, Claude) de kabul edilebilir düzeyde klinik destek sağlayabilir. İkinci olarak, halüsinasyon riski ve kaynak izlenebilirliği kritik öneme sahip durumlarda RAG tabanlı sistemler avantajlı görünmektedir. Üçüncü olarak, pediatrik gastroenteroloji gibi özel alanlarda bilgi tabanı genişletilmesi veya genel amaçlı model kullanımı değerlendirilmelidir (88).

Yapay zeka destekli klinik karar destek sistemlerinin gastroenteroloji pratiğine entegrasyonu aşamalı bir süreç olmalıdır (89). Başlangıç aşamasında düşük riskli görevler (hasta eğitim materyali üretimi, randevu öncesi semptom değerlendirmesi, rutin takip önerileri) için kullanım uygun olabilir. İleri aşamalarda, klinik validasyon çalışmalarının ardından tanısal karar desteği ve tedavi önerisi sistemleri devreye

alınabilir. Ancak kritik kararlar (kanser tanısı, cerrahi endikasyon, ileri tedavi seçenekleri) için yapay zeka çıktıları her zaman uzman klinisyen tarafından doğrulanmalıdır (90).

Düzenleyici çerçeve açısından, FDA'nın Software as a Medical Device (SaMD) sınıflandırması ve CE işaretleme gereklilikleri, tıbbi yapay zeka uygulamalarının piyasaya sürülmesinde belirleyici olmaktadır. Sınıf II (orta risk) veya Sınıf III (yüksek risk) tıbbi cihaz olarak sınıflandırılan yapay zeka sistemleri, kapsamlı klinik validasyon ve güvenlik değerlendirmesi gerektirmektedir .Çalışmamız, bu validasyon süreçleri için metodolojik bir çerçeve sunmaktadır .(91)

Etik boyut açısından, yapay zeka modellerinin tıbbi kararlardaki rolü şeffaf olmalı ve hasta bilgilendirilmiş onayı kapsamında değerlendirilmelidir. Algoritmik önyargı (bias), veri gizliliği ve hesap verebilirlik (accountability) konuları, yapay zeka destekli sağlık hizmetlerinin etik çerçevesini belirlemektedir (92). Modellerin eğitim verilerindeki demografik veya coğrafi dengesizlikler, belirli hasta gruplarında performans farklılıklarına yol açabilir ve sağlık eşitsizliklerini derinleştirebilir (93).

Çalışmamızın önemli güçlü yönleri bulunmaktadır. İlk olarak, 200 standardize klinik vaka içeren kapsamlı veri seti, gastroenterolojinin farklı alt alanlarını ve karmaşıklık düzeylerini temsil etmektedir (94). İkinci olarak, 8 farklı modelin sistematik karşılaştırması, literatürdeki en kapsamlı çalışmalardan birini oluşturmaktadır (95). Üçüncü olarak, körleştirilmiş değerlendirme ve iki bağımsız uzman değerlendiricinin kullanımı, sonuçların objektifliğini güçlendirmiştir (96). Dördüncü olarak, çok boyutlu değerlendirme kriterleri (tanısal doğruluk, ayırıcı tanı, tetkik önerileri, tedavi planı, hasta danışmanlığı, genel muhakeme), klinik performansın bütüncül değerlendirmesini sağlamıştır (97). Beşinci olarak, halüsinasyon, tutarlılık ve kaynak atfı gibi güvenlik parametrelerinin değerlendirilmesi, klinik uygulama açısından kritik bilgiler sunmuştur (98) .

Çalışmamızın bazı sınırlılıkları mevcuttur (99). İlk olarak, GastroGPT-Faiss (n=45) ve GastroGPT-Chroma (n=12) için düşük yanıt oranları, bu modellerin değerlendirmesinde istatistiksel gücü sınırlamıştır. İkinci olarak, vakalar İngilizce olarak sunulmuş olup, Türkçe veya diğer dillerdeki performans farklılıkları

değerlendirilmemiştir (100). Üçüncü olarak, gerçek klinik ortam simülasyonu yerine standardize vakalar kullanılmış olup, gerçek hasta etkileşimlerindeki performans farklılık gösterebilir (101). Dördüncü olarak, model versiyonları çalışma süresince sabitlenmiş olup, sürekli güncellenen modellerin (GPT-4 Turbo, Claude 3 vb.) performansı değerlendirilmemiştir. Beşinci olarak, maliyet-etkinlik analizi çalışma kapsamına dahil edilmemiştir. Altıncı olarak, uzun dönemli hasta sonlanımları değerlendirilmemiş olup, yapay zeka destekli kararların klinik sonuçlara etkisi bilinmemektedir (102).

Gelecek araştırmalar için şu öneriler sunulmaktadır(103) . İlk olarak, prospektif klinik çalışmalar ile yapay zeka destekli karar destek sistemlerinin hasta sonlanımlarına etkisi değerlendirilmelidir (104). İkinci olarak, çok merkezli ve çok uluslu çalışmalar ile modellerin farklı sağlık sistemleri ve hasta popülasyonlarındaki performansı karşılaştırılmalıdır. Üçüncü olarak, Türkçe ve diğer dillerde model performansının değerlendirilmesi, yerel uygulamalar için kritik önem taşımaktadır (105). Dördüncü olarak, pediatrik, geriatrik ve gebe hasta grupları gibi özel popülasyonlara özgü model geliştirme ve validasyon çalışmaları yürütülmelidir (106). Beşinci olarak, görüntü-metin multimodal modellerin endoskopik görüntü yorumlama ve SOAP notu üretimindeki performansı değerlendirilmelidir (107). Altıncı olarak, gerçek zamanlı klinik entegrasyon çalışmaları ile kullanıcı deneyimi ve iş akışı etkileri araştırılmalıdır. Yedinci olarak, maliyet-etkinlik analizleri ile yapay zeka destekli sistemlerin sağlık ekonomisi üzerindeki etkisi değerlendirilmelidir (108).

Mevcut çalışma gastroenterolojiye özgü bir yapay zeka dil modelinin (GastroGPT) klinik karar destek kapasitesini kapsamlı bir şekilde değerlendirmiş ve genel amaçlı modellerle karşılaştırmalı analizini gerçekleştirmiştir (109). Bulgularımız, GastroGPT-Deeplake'in genel performansta GPT-4 ve Claude'a göre üstünlüğünü ve aynı zamanda halüsinasyon oranı ile kaynak atfı kalitesi açısından da belirgin avantajlar sunduğunu ortaya koymuştur (110). Alt uzmanlık alanlarına göre performans farklılıkları, model seçiminin klinik bağlama göre optimize edilmesi gerektiğini göstermektedir (111). Vaka karmaşıklığı arttıkça alan-spesifik bilgi tabanının belirleyici hale gelmesi, GastroGPT gibi RAG tabanlı modellerin klinik karar desteğindeki kritik rolünü desteklemektedir (112). RAG entegrasyonu, tıbbi

yapay zeka uygulamalarında güvenlik ve izlenebilirlik açısından kritik bir bileşen olarak değerlendirilmelidir (113). Yapay zeka destekli klinik karar destek sistemlerinin gastroenteroloji pratiğine güvenli ve etkin şekilde entegrasyonu, süregelen araştırma, validasyon ve düzenleyici uyum çalışmalarını gerektirmektedir (114).

Multimodal yapay zeka sistemlerinin gastroenterolojideki potansiyeli de tartışılmalıdır (115). Endoskopik görüntü analizi ile metin tabanlı klinik muhakemeyi birleştiren sistemler, gelecekte tanısal doğruluğu artırabilir. GPT-4V ve benzeri görsel-dil modellerinin endoskopi raporlaması ile entegrasyonu, daha kapsamlı klinik karar destek sistemlerinin geliştirilmesine olanak tanıyacaktır (116). Yapay zeka modellerinin sürekli öğrenme (continual learning) kapasitesi, tıbbi uygulamalarda önemli bir konu olarak öne çıkmaktadır (117). Güncel rehberler, yeni ilaç onayları ve değişen klinik pratikler, modellerin düzenli güncellenmesini gerektirmektedir. RAG sistemleri, bilgi tabanının güncellenmesi yoluyla bu soruna kısmi çözüm sunmakta; ancak temel modelin de periyodik olarak yeniden eğitilmesi veya fine-tuning işlemine tabi tutulması gerekebilir (118). Yapay zeka modellerinin açıklanabilirliği (explainability) ve yorumlanabilirliği (interpretability), tıbbi uygulamalarda kritik öneme sahiptir (119). Klinisyenler, model önerilerinin arkasındaki mantığı anlayabilmeli ve gerektiğinde sorgulayabilmelidir. GastroGPT'nin kaynak atfı özelliği, bu açıdan önemli bir adım olmakla birlikte, modelin muhakeme sürecinin tam şeffaflığı hâlâ sağlanamamaktadır (120). Farklı hasta popülasyonlarında model performansının değişkenliği, algoritmik önyargı (bias) açısından değerlendirilmelidir (121). Modellerin eğitim verileri ağırlıklı olarak Batı toplumlarından ve akademik merkezlerden gelmekte olup, farklı etnik gruplar, coğrafi bölgeler ve sosyoekonomik düzeylerdeki hastalarda performans farklılıkları görülebilir (122). Türkiye özgü hastalık spektrumu (Behçet hastalığı, Akdeniz ateşi, hepatit B prevalansı vb.) göz önünde bulundurularak yerelleştirilmiş modellerin geliştirilmesi önerilmektedir (123).

Sağlık ekonomisi perspektifinden, yapay zeka destekli sistemlerin maliyet-etkinliği kapsamlı değerlendirilmelidir (124). Model kullanım maliyetleri (API ücretleri, hesaplama kaynakları), entegrasyon maliyetleri, eğitim gereksinimleri ve potansiyel tasarruflar (tanısal doğruluk artışı, gereksiz tetkik azalması, komplikasyon

önleme) birlikte ele alınmalıdır. Gelecek çalışmalarda formal maliyet-etkinlik analizlerinin yapılması önerilmektedir (125). Klinisyen-yapay zeka etkileşimi dinamikleri, sistemlerin klinik pratiğe entegrasyonunda belirleyici rol oynamaktadır (126). Otomasyon önyargısı (automation bias), klinisyenlerin model önerilerini eleştirmeden kabul etme eğilimini ifade etmekte ve potansiyel riskler taşımaktadır (127). Öte yandan aşırı şüphecilik, yapay zeka sistemlerinin potansiyel faydalarından yararlanılmasını engelleyebilir (128). Dengeli bir yaklaşım için klinisyen eğitimi ve kullanıcı arayüzü tasarımı kritik öneme sahiptir (129).

Hukuki sorumluluk (liability) konusu, tıbbi yapay zeka uygulamalarının yaygınlaşmasıyla giderek önem kazanmaktadır (130).Yapay zeka destekli bir kararda hata oluştuğunda sorumluluğun kime ait olacağı (klinisyen, hastane, model geliştiricisi, yazılım sağlayıcısı) açık değildir (131). Düzenleyici çerçevelerin ve sigorta sistemlerinin bu belirsizliği giderecek şekilde güncellenmesi gerekmektedir (132). Hasta perspektifi ve kabul edilebilirlik, yapay zeka destekli sağlık hizmetlerinin başarısında önemli bir faktördür (133) .Hastaların yapay zeka kullanımı hakkında bilgilendirilmesi, onay alınması ve endişelerinin giderilmesi, etik ve yasal açıdan gereklidir (134) .Çalışmalar, hastaların yapay zeka destekli sistemlere karşı tutumlarının değişken olduğunu ve eğitim düzeyi, yaş, sağlık okuryazarlığı gibi faktörlerden etkilendiğini göstermektedir (135).

Derin öğrenme modellerinin gastroenterolojideki klinik uygulamaları, görüntü analizi ile sınırlı kalmayıp metin tabanlı karar desteğine doğru genişlemektedir (135) .Bu çalışma, bu geçişin öncü örneklerinden birini sunmakta ve alan-spesifik ile genel amaçlı modellerin karşılaştırmalı değerlendirmesini sağlamaktadır (136).Bulgularımız, her iki yaklaşımın da kendine özgü avantajları olduğunu ve hibrit sistemlerin gelecekte en iyi sonuçları verebileceğini düşündürmektedir(137). Yapay zeka modellerinin sağlık hizmetlerinde kullanımı, veri kalitesi ve standardizasyonu konusunda önemli zorluklar barındırmaktadır. Farklı elektronik sağlık kaydı sistemleri, kodlama standartları ve dokümantasyon pratikleri, modellerin gerçek dünya performansını etkileyebilir (138). HL7 FHIR gibi interoperabilite standartlarının yaygınlaşması, bu sorunların çözümüne katkı sağlayabilir (138, 139). Federe öğrenme (federated learning) yaklaşımları, hasta verilerinin merkezi toplanması gerektirmeden

çok merkezli model geliştirmeye olanak tanımaktadır (140). Bu yaklaşım, veri gizliliği endişelerini giderirken çeşitli hasta popülasyonlarından öğrenmeyi mümkün kılmaktadır (141). Gastroenteroloji alanında federe öğrenme ile geliştirilen modellerin performansının değerlendirilmesi, gelecek araştırma konuları arasında yer almaktadır (142).

Modellerin kalibrasyon kalitesi, yani tahmin güvenilirliğinin gerçek doğruluk ile uyumu, klinik uygulamalarda kritik öneme sahiptir (143). Aşırı güvenli (overconfident) modeller, hatalı önerileri bile yüksek güvenle sunabilir ve klinisyeni yanıltabilir. Belirsizlik tahmini (uncertainty quantification) tekniklerinin entegrasyonu, model çıktılarının güvenilirliğini artırabilir (144). Prompt engineering, yapay zeka modellerinden optimal çıktı elde etmede belirleyici bir faktördür (145) .Çalışmamızda kullanılan standart SOAP notu promptu, tüm modeller için eşit koşullar sağlamış olmakla birlikte, model-spesifik prompt optimizasyonunun performansı artırıp artırmayacağı araştırılmamıştır (146). Chain-of-thought, few-shot learning ve role prompting gibi ileri prompt teknikleri, gelecek çalışmalarda değerlendirilebilir (147) .

Yapay zeka modellerinin real-time klinik entegrasyonu, teknik ve organizasyonel zorluklar içermektedir (148). Sistem yanıt süreleri, kullanıcı arayüzü tasarımı, mevcut iş akışlarıyla uyum ve IT altyapısı gereksinimleri, başarılı implementasyonun belirleyicileri arasındadır (149). Çalışmamızda tüm modellerin yanıt süreleri klinik kullanım için kabul edilebilir sınırlar içinde (<10 saniye) bulunmuştur (150). Eğitim ve kapasite geliştirme, yapay zeka destekli sistemlerin klinik pratiğe entegrasyonunun kritik bileşenidir (151). Klinisyenlerin yapay zeka okuryazarlığı, model kapasiteleri ve sınırlılıkları hakkında bilgilendirilmesi, sistemlerin güvenli ve etkin kullanımı için şarttır (152).Tıp fakültesi müfredatlarında yapay zeka eğitiminin yer alması, geleceğin hekimlerini bu dönüşüme hazırlayacaktır (153).

Farklı sağlık sistemleri ve ödeme modelleri, yapay zeka uygulamalarının benimsenmesini etkilemektedir (154) .Fee-for-service sistemlerde verimlilik artışı ekonomik teşvik sağlarken, capitation modellerinde maliyet azaltımı ön plana

çıkmaktadır .Türkiye sağlık sistemi bağlamında, SGK geri ödeme politikalarının yapay zeka destekli hizmetleri nasıl karşılayacağı belirsizliğini korumaktadır (155).

Klinik karar destek sistemlerinin uyarı yorgunluğu (alert fatigue) riski, gastroenteroloji pratiğinde de geçerlidir (156). Aşırı sayıda veya düşük özgüllükte uyarılar, klinisyenlerin sistemi görmezden gelmesine yol açabilir (157). Model çıktılarının klinik bağlama göre filtrelenmesi ve önceliklendirilmesi, bu riski azaltabilir (158).

Yapay zeka modellerinin güncelleme ve bakım döngüsü, klinik uygulamalarda sürdürülebilirlik açısından önemlidir (159). Model degradasyonu (model drift), zaman içinde değişen veri dağılımları nedeniyle performans düşüşüne yol açabilir (160) .Sürekli izleme, periyodik değerlendirme ve gerektiğinde yeniden eğitim, sistemlerin uzun vadeli etkinliği için gereklidir (161). Türkiye'de gastroenteroloji pratiğinin özgün koşulları, yapay zeka uygulamalarının yerelleştirilmesini gerektirmektedir. Hastalık spektrumu (viral hepatit B prevalansı, Akdeniz tipi G6PD eksikliği, Behçet hastalığı vb.), sağlık altyapısı ve kaynak kısıtlamaları, Batı ülkelerinden farklılık göstermektedir. Yerel validasyon çalışmaları ve gerektiğinde model adaptasyonu, Türkiye'de güvenli uygulama için şarttır (161).

Multidisipliner işbirliği, tıbbi yapay zeka sistemlerinin başarılı geliştirilmesi ve uygulanmasında vazgeçilmezdir (162). Gastroenterologlar, veri bilimciler, yazılım mühendisleri, etik uzmanları ve sağlık yöneticilerinin ortak çalışması, kapsamlı ve sürdürülebilir çözümler üretebilir (163). Bu çalışma, bu tür interdisipliner işbirliğinin somut bir örneğini oluşturmaktadır (164).

Gelecekte yapay zeka sistemlerinin gastroenteroloji pratiğindeki rolü muhtemelen genişleyecektir (165). Tanısal desteğin ötesinde, tedavi yanıtı tahmini, hastalık prognozu, personalize tıp uygulamaları ve popülasyon sağlığı yönetimi alanlarında kullanım potansiyeli mevcuttur (166). Bu genişleme, süregelen araştırma, validasyon ve etik tartışmalarla desteklenmelidir (167). Yapay zeka ve insan zekasının tamamlayıcı güçleri, optimal klinik sonuçlar için birleştirilmelidir (168). Yapay zekanın hızı, tutarlılığı ve geniş veri işleme kapasitesi; insan hekimin empati, bağlamsal yargı ve karmaşık etik değerlendirme yetenekleri ile birleştiğinde, hasta

bakımında en iyi sonuçlar elde edilebilir (169). Bu sinerji, gelecekte sağlık hizmetlerinin temel paradigmasını oluşturacaktır (170) .

Yapay zeka modellerinin gastroenteroloji alanındaki klinik uygulama potansiyeli, bu çalışmanın bulgularıyla somut verilerle desteklenmiştir (171). Hem alan-spesifik hem de genel amaçlı modellerin güçlü ve zayıf yönlerinin sistematik olarak değerlendirilmesi, gelecek araştırma ve uygulama stratejilerinin belirlenmesine katkı sağlamaktadır (172).

Model performansının değerlendirmesinde kullanılan SOAP notu formatı, klinik muhakemenin çok boyutlu değerlendirmesine olanak tanımıştır (173). Geleneksel çoktan seçmeli sınav formatlarının aksine, SOAP değerlendirmesi tanısal düşünme, ayırıcı tanı oluşturma, tetkik planlama, tedavi kararı ve hasta iletişimi gibi bileşenlerin ayrı ayrı skorlanmasını sağlamıştır (174). Bu yaklaşım, modellerin klinik muhakeme sürecindeki spesifik güçlü ve zayıf yönlerinin belirlenmesine imkan vermiştir (175).

Gastroenterolojinin heterojen doğası, yapay zeka modellerinin tek tip bir performans göstermesini zorlaştırmaktadır (176). Üst gastrointestinal sistem patolojileri, alt gastrointestinal sistem hastalıkları, hepatobiliyer sistem, pankreas ve ince bağırsak hastalıkları farklı tanısal ve terapötik yaklaşımlar gerektirmektedir (177). Bu çeşitlilik, alan-spesifik modellerin bile tüm alt alanlarda eşit düzeyde performans göstermesini güçleştirmektedir (178). Endoskopik prosedürlerin yapay zeka ile desteklenmesi, gastroenterolojide en gelişmiş uygulama alanını oluşturmaktadır (179). Bu çalışmada değerlendirilen metin tabanlı modeller, endoskopik görüntü analizi yerine klinik muhakeme süreçlerine odaklanmıştır (180). Gelecekte görüntü ve metin tabanlı modellerin entegrasyonu, kapsamlı klinik karar destek sistemlerinin oluşturulmasına olanak tanıyacaktır (181).

Yapay zeka modellerinin eğitim ve sürekli mesleki gelişim programlarındaki rolü de değerlendirilmelidir (182). Simüle vaka çalışmaları, ayırıcı tanı egzersizleri ve tedavi algoritması öğrenimi için bu modeller değerli eğitim araçları olabilir (183). Tıp öğrencileri ve asistan hekimlerin yapay zeka destekli öğrenme ortamlarından faydalanması, gelecek nesil klinisyenlerin yetiştirilmesine katkı sağlayabilir (184).

Hasta güvenliği perspektifinden, yapay zeka modellerinin potansiyel riskleri sistematik olarak yönetilmelidir (185). Yanlış pozitif ve yanlış negatif oranları, modelin kullanım senaryosuna göre farklı riskler taşımaktadır. Tarama amaçlı kullanımda yanlış pozitifler gereksiz anksiyete ve tetkiklere, tedavi önerilerinde yanlış negatifler ciddi hastalıkların atlanmasına yol açabilir (186).

Sağlık hizmeti sunucuları ve yapay zeka geliştiricileri arasındaki işbirliği, başarılı implementasyonun anahtarıdır (187). Klinisyenlerin ihtiyaçlarının anlaşılması, kullanılabilir arayüzlerin tasarlanması ve gerçek dünya geri bildirimlerine dayalı iyileştirmeler, sistemlerin benimsenmesini artıracaktır (188).

Yapay zeka modellerinin performansının uzun vadeli izlenmesi, klinik uygulamanın sürdürülebilirliği için kritiktir (189). Model degradasyonu, değişen klinik pratikler ve yeni kanıtların ortaya çıkması, periyodik değerlendirme ve güncelleme gerektirmektedir (190) Sürekli performans izleme sistemlerinin kurulması, güvenli ve etkin kullanımın sağlanması için vazgeçilmezdir (191).

Çalışmamızın metodolojik yaklaşımı, gelecek araştırmalar için bir çerçeve sunmaktadır (192). Standardize vaka setleri, körleştirilmiş değerlendirme, çoklu değerlendirici kullanımı ve kapsamlı istatistiksel analiz, tıbbi yapay zeka çalışmalarının kalitesini artıran unsurlar olarak önerilmektedir (193). Bu metodolojik standartların yaygınlaşması, çalışmalar arası karşılaştırılabilirliği ve kanıt sentezini kolaylaştıracaktır (194).

Sonuç olarak, bu çalışma gastroenterolojiye özgü yapay zeka dil modellerinin klinik potansiyelini ve sınırlılıklarını kapsamlı olarak ortaya koymuştur (195). RAG entegrasyonunun halüsinasyon azaltma ve kaynak izlenebilirlik açısından sağladığı avantajlar, tıbbi yapay zeka uygulamalarının güvenliği için kritik öneme sahiptir (196). GastroGPT-Deeplake'in hem genel hem de karmaşık vakalardaki üstünlüğü, alan-spesifik RAG entegrasyonunun gastroenteroloji pratiğinde en etkili yaklaşım olduğunu göstermektedir (196). Yapay zeka destekli klinik karar destek sistemlerinin gastroenteroloji pratiğine başarılı entegrasyonu, multidisipliner işbirliği, sürekli araştırma ve düzenleyici uyum çalışmalarını gerektirmektedir (197).

İnflamatuvar bağırsak hastalıklarının yönetiminde yapay zeka modellerinin rolü, özellikle tedavi yanıtı tahmini ve hastalık aktivitesi değerlendirmesinde genişleme potansiyeli taşımaktadır (198). Crohn hastalığı ve ülseratif kolit hastalarının heterojen tedavi yanıtları, kişiselleştirilmiş tedavi yaklaşımlarını zorunlu kılmakta ve yapay zeka bu bireyselleştirme sürecine katkı sağlayabilir (199). İBH'da kullanılan endoskopik skorlama sistemlerinin (Mayo, SES-CD, UCEIS vb.) standardizasyonu ve otomatik değerlendirmesi, yapay zeka destekli sistemler ile daha tutarlı hale gelebilir (200).

Karaciğer hastalıklarında non-invaziv fibrozis değerlendirmesi, yapay zeka uygulamalarının en umut verici alanlarından birini oluşturmaktadır (201). FIB-4, APRI gibi skorların yanı sıra elastografi ve biyobelirteç kombinasyonlarının yapay zeka ile entegrasyonu, karaciğer biyopsisi ihtiyacını azaltabilir (202). GastroGPT'nin hepatoloji vakalarındaki göreceli başarısı, bu alandaki potansiyeli desteklemektedir (203).

Pediatrik gastroenteroloji alanında yapay zeka uygulamalarının geliştirilmesi, özel zorluklar içermektedir (204). Çocuklarda hastalık prezentasyonlarının farklılığı, yaşa özgü normal değerler, büyüme ve gelişim parametreleri ile pediatrik ilaç dozajları, erişkin modellerinden farklı yaklaşımlar gerektirmektedir (205). NASPGHAN ve ESPGHAN rehberlerinin bilgi tabanına eklenmesi ve pediatrik vaka setleri ile ince ayar yapılması, bu alandaki performansı artırabilir (206) .

Gastrointestinal onkoloji alanında yapay zeka, erken tanıdan palyatif bakıma kadar geniş bir spektrumda destek sağlayabilir (46). Kolorektal kanser taramasında adenom tespit oranının artırılması, Barrett özofagus sürveyansı, gastrik kanser risk değerlendirmesi ve kemoterapi rejimi seçimi, potansiyel uygulama alanları arasındadır. Multidisipliner tümör kurulu tartışmalarının yapay zeka ile desteklenmesi, karar kalitesini artırabilir (207).

Fonksiyonel gastrointestinal bozukluklar, tanısal belirsizlik nedeniyle yapay zeka uygulamalarının zorlandığı bir alan olarak öne çıkmaktadır (208). Roma IV kriterlerinin subjektif değerlendirme bileşenleri ve organik patolojinin ekartasyonu gerekliliği, algoritmik yaklaşımları karmaşıklaştırmaktadır (209). Bununla birlikte,

semptom örüntüsü tanıma ve tedavi yanıtı tahmini alanlarında yapay zeka potansiyel taşımaktadır (210) .

Acil gastroenterolojik durumların yönetiminde yapay zeka, triaj ve erken müdahale kararlarını destekleyebilir (211). Akut gastrointestinal kanama, akut pankreatit, akut kolesistit ve bağırsak obstrüksiyonu gibi durumların şiddet değerlendirmesi ve yönetim önceliklendirmesi, yapay zeka destekli sistemlerle optimize edilebilir (212). GastroGPT'nin pankreatit şiddet skorlamasındaki başarısı, bu potansiyeli desteklemektedir (213).

Nutrisyonel değerlendirme ve malnütrisyon yönetimi, gastroenteroloji pratiğinin önemli bir bileşenidir (214). Yapay zeka modelleri, beslenme durumu değerlendirmesi, enteral/parenteral nütrisyon kararları ve diyet önerilerinde destek sağlayabilir (215). Özellikle kısa bağırsak sendromu, pankreas yetmezliği ve malabsorbsiyon durumlarında bireyselleştirilmiş beslenme planlaması kritik öneme sahiptir (216).

Gastroenterolojik girişimsel prosedürlerin planlanması ve risk değerlendirmesinde yapay zeka, pre-prosedürel değerlendirmeyi standardize edebilir (217). ERCP, endoskopik ultrasonografi, endoskopik mukozal rezeksiyon ve peroral endoskopik miyotomi gibi işlemlerin hasta seçimi ve komplikasyon tahmini, yapay zeka destekli sistemlerle optimize edilebilir (218).

Kronik karaciğer hastalığının komplikasyonlarının yönetiminde yapay zeka, hepatik ensefalopati, asit, hepatorenal sendrom ve variseal kanama risk değerlendirmesinde destek sağlayabilir (219). MELD skoru optimizasyonu ve transplantasyon önceliklendirmesi, yapay zeka uygulamalarının potansiyel katkı sağlayabileceği kritik alanlardır (220).

Gastrointestinal motilite bozukluklarının tanı ve yönetiminde yapay zeka, manometri verilerinin yorumlanması ve tedavi yanıtı tahminine katkı sağlayabilir (221). Özofagus akalazya, gastroparezi ve kronik intestinal psödo-obstrüksiyon gibi durumların kompleks patofizyolojisi, yapay zeka destekli karar desteğinden faydalanabilir (222).

Yapay zeka modellerinin güncelleme stratejileri, klinik uygulamanın sürdürülebilirliği için kritiktir (223). Tıbbi bilginin hızla değiştiği günümüzde, modellerin güncel kalması için sistematik güncelleme protokolleri gereklidir (224). RAG sistemlerinin bilgi tabanı güncellemeleri, temel model güncellemeye göre daha pratik bir çözüm sunmaktadır (225).

Hasta-hekim ilişkisinin yapay zeka entegrasyonundan nasıl etkileneceği, önemli bir araştırma ve tartışma konusudur (226). Yapay zeka destekli sistemlerin klinik karşılaşmada nasıl konumlandırılacağı, hasta güveni ve memnuniyetini etkileyebilir. Şeffaf iletişim ve yapay zekanın destekleyici rolünün vurgulanması, bu endişelerin giderilmesine katkı sağlayabilir (227).

Yapay zeka sistemlerinin denetim ve kalite güvencesi, sağlık hizmetlerinin güvenliği için vazgeçilmezdir (228). Model performansının sürekli izlenmesi, hata raporlama sistemleri ve periyodik kalibrasyon, güvenli kullanımın temel bileşenleridir. Bu çalışmada kullanılan değerlendirme metodolojisi, kalite güvence çerçevesi için bir model sunmaktadır (229).

Gelecekte yapay zeka ve insan zekasının optimal entegrasyonu, gastroenteroloji pratiğinin evriminde belirleyici olacaktır (230). Yapay zekanın rutin ve tekrarlayan görevlerde verimliliği artırması, klinisyenlerin karmaşık karar verme ve hasta ilişkilerine daha fazla zaman ayırmasına olanak tanıyabilir (231). Bu sinerji, hasta bakımının kalitesini ve klinisyen iş tatminini artırma potansiyeli taşımaktadır (232).

Yapay zeka modellerinin standardizasyonu ve interoperabilitesi, farklı sağlık sistemleri arasında tutarlı uygulamayı mümkün kılacaktır. HL7 FHIR ve SMART on FHIR gibi standartların yapay zeka sistemleri ile entegrasyonu, yaygın benimseme için kritik öneme sahiptir (233).

Klinik karar destek sistemlerinin tasarımında kullanıcı deneyimi (UX) ilkeleri, sistemlerin benimsenmesini ve etkin kullanımını belirlemektedir (234). Gastroenterologların iş akışına entegre, minimal bilişsel yük oluşturan ve zamanında geri bildirim sağlayan arayüzler, başarılı implementasyonun temel bileşenleridir (235).

Yapay zeka sistemlerinin klinik performansının gerçek dünya verileri ile validasyonu, kontrollü çalışma sonuçlarından farklılık gösterebilir. Gerçek klinik ortamın karmaşıklığı, veri kalitesi sorunları ve kullanıcı davranış değişkenlikleri, model performansını etkileyebilir (236). Yapay zeka destekli gastroenteroloji uygulamalarının global erişilebilirliği, sağlık eşitliği perspektifinden değerlendirilmelidir (237). Düşük ve orta gelirli ülkelerde uzman hekim yetersizliği, yapay zeka destekli sistemlerin değerini artırmakta; ancak teknoloji altyapısı ve maliyet engelleri erişimi sınırlayabilmektedir (238).

Sonuç olarak, bu kapsamlı çalışma gastroenterolojiye özgü yapay zeka dil modellerinin klinik potansiyelini, sınırlılıklarını ve gelecek yönelimlerini detaylı olarak ortaya koymuştur (239). Bulgularımız, alan-spesifik RAG sistemlerinin güvenlik avantajları ile genel amaçlı modellerin muhakeme kapasitesinin birleştirilmesi gerektiğini göstermektedir (240). Yapay zeka destekli klinik karar destek sistemlerinin gastroenteroloji pratiğine başarılı entegrasyonu, multidisipliner işbirliği, sürekli araştırma, etik değerlendirme ve düzenleyici uyum çalışmalarının kesişim noktasında şekillenecektir (241).

Yapay zeka modellerinin klinik karar süreçlerinde kullanımı, tıbbi eğitim paradigmalarının da güncellenmesini gerektirmektedir (242). Gelecek nesil gastroenterologların yapay zeka okuryazarlığı kazanması, bu sistemlerin etkin ve eleştirel kullanımı için şarttır (243) . Tıp fakültelerinde ve uzmanlık eğitimi programlarında yapay zeka müfredatının yer alması önerilmektedir (244).

Yapay zeka sistemlerinin gastroenterolojideki uygulama alanları, klinik pratiğin ötesine geçerek araştırma metodolojisini de dönüştürme potansiyeli taşımaktadır (245). Büyük veri analitiği, kohort tanımlama, outcome prediction ve randomize kontrollü çalışma tasarımı optimizasyonu, yapay zekanın katkı sağlayabileceği araştırma alanlarıdır (246).

Gastroenteroloji pratiğinde yapay zeka uygulamalarının etik boyutları, sürekli tartışma ve değerlendirme gerektirmektedir (247). Algoritmik adalet, veri mahremiyeti, bilgilendirilmiş onam, hesap verebilirlik ve insan-makine etkileşiminin sınırları, çözülmesi gereken temel etik meseleler arasındadır (248).

Bu çalışmanın bulguları, gastroenterolojide yapay zeka dil modellerinin klinik entegrasyonu için somut bir yol haritası sunmaktadır (249). RAG tabanlı sistemlerin güvenlik avantajları, genel amaçlı modellerin muhakeme kapasitesi ve klinisyen uzmanlığının birleştirilmesi, optimal hasta bakımı için gerekli sinerjiyi oluşturacaktır (250) .Yapay zeka, gastroenteroloji pratiğini dönüştürme potansiyeli taşımakta olup, bu dönüşümün güvenli, etik ve hasta merkezli olması için multidisipliner çabaların sürdürülmesi gerekmektedir (251) .

Yapay zeka sistemlerinin gastroenteroloji pratiğindeki uzun vadeli etkilerinin izlenmesi, sistematik veri toplama ve analiz gerektirmektedir(252). Klinik sonuçlar, hasta memnuniyeti, klinisyen iş yükü ve sağlık hizmeti maliyetleri üzerindeki etkilerin prospektif değerlendirilmesi, kanıt tabanının güçlendirilmesi için şarttır (253). Bu tür uzun dönemli izlem çalışmaları, yapay zekâ uygulamalarının gerçek dünya değerini ortaya koyacaktır (254).

Yapay zeka modellerinin güvenlik izlemi, farmakovijilans benzeri sistemlerin oluşturulmasını gerektirmektedir. Advers olay raporlama, performans sapma tespiti ve düzeltici eylem mekanizmalarının kurulması, hasta güvenliğinin korunması için kritik öneme sahiptir.

GASTROGPT klinik performans açısından mevcut yapay zeka dil modellerine kıyasla üstün performans göstermiştir. Mevcut analiz, gittikçe kullanımı yaygınlaşan yapay zeka dil modellerinin branşa spesifik dil modellerinin önemi vurgulamaktadır. Yapay zeka modelleri hasta yönetiminde klinisyene yardımcı modellerdir. Klinisyen, tanısal yaklaşım ve hasta yönetimini yapabilecek yetkinliğe sahiptir. Yüksek karmaşıklığa sahip vakalarda yapay zeka modellerinin iki aşamalı yaklaşım metodu ile kullanımı özellikle karmaşık vakalarda uygun tetkik korelasyonunu arttırabilir. İlk aşamada tanısal ön tanıların daraltılması, ikinci aşamada daraltılmış ön tanılara yönelik hasta yönetimi, yapay zeka modellerinin hasta yönetiminde yüksek korelasyonla çalışmalarına olanak sağlayabilir.

5. SONUÇLAR

1. GastroGPT-Deeplake (53.8±5.9) genel performansta en yüksek puanı elde etmiş olup, GPT-4 (50.1±6.8) ve Claude (49.3±7.2) modellerini anlamlı düzeyde geride bırakmıştır.

2. GastroGPT varyantları, halüsinasyon oranları (%4.2-6.8 vs %8.3-12.1, p=0.018) ve kaynak atfı kalitesi (%62.8-78.3 vs %23.5-34.7) açısından belirgin üstünlük göstermiştir.

3. Hepatoloji vakalarında GastroGPT-Deeplake, GPT-4'den anlamlı düzeyde üstün performans göstermiştir (p=0.034). Tüm alt uzmanlık alanlarında GastroGPT-Deeplake tutarlı üstünlük sergilemiştir.

4. Vaka karmaşıklığı arttıkça GastroGPT-Deeplake'in alan-spesifik avantajı belirginleşmiş; yüksek karmaşıklıkta GastroGPT-Deeplake anlamlı üstünlük göstermiştir ($p<0.001$).

5. RAG entegrasyonu, halüsinasyon riskini azaltmada ve kaynak izlenebilirliğini sağlamada kritik bir bileşen olarak değerlendirilmelidir.

6. Yapay zeka destekli sistemlerin gastroenteroloji pratiğine güvenli entegrasyonu için süregelen araştırma ve validasyon çalışmaları gerekmektedir.

KAYNAKÇA

1. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine. 2019;25(1):44-56.

2. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med. 2022;28(1):31-8.

3. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nature Medicine. 2019;25(1):24-9.

4. Char DS, Shah NH, Magnus D. Implementing Machine Learning in Health Care - Addressing Ethical Challenges. N Engl J Med. 2018;378(11):981-3.

5. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436-44.

6. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Communications of the ACM. 2012;60:84 - 90.

7. Wu J, Chen J, Cai J. Application of Artificial Intelligence in Gastrointestinal Endoscopy. J Clin Gastroenterol. 2021;55(2):110-20.

8. Le Berre C, Sandborn WJ, Aridhi S, Devignes MD, Fournier L, Smaïl-Tabbone M, et al. Application of Artificial Intelligence to Gastroenterology and Hepatology. Gastroenterology. 2020;158(1):76-94.e2.

9. Arnold M, Abnet CC, Neale RE, Vignat J, Giovannucci EL, McGlynn KA, et al. Global Burden of 5 Major Types of Gastrointestinal Cancer. Gastroenterology. 2020;159(1):335-49.e15.

10. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. 2021;71(3):209-49.

11. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. Proceedings of the 34th International Conference on Neural Information Processing Systems; Vancouver, BC, Canada: Curran Associates Inc.; 2020. p. Article 159.

12. Wiggins WF, Tejani AS. On the Opportunities and Risks of Foundation Models for Natural Language Processing in Radiology. Radiol Artif Intell. 2022;4(4):e220119.

13. Devlin J, Chang M-W, Lee K, Toutanova K, editors. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. North American Chapter of the Association for Computational Linguistics; 2019.

14. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023;29(8):1930-40.

15. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2(2):e0000198.

16. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks2020.

17. Wang P, Berzin TM, Glissen Brown JR, Bharadwaj S, Becq A, Xiao X, et al. Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. Gut. 2019;68(10):1813-9.

18. Spadaccini M, Iannone A, Maselli R, Badalamenti M, Desai M, Chandrasekar VT, et al. Computer-aided detection versus advanced imaging for detection of colorectal neoplasia: a systematic review and network meta-analysis. Lancet Gastroenterol Hepatol. 2021;6(10):793-802.

19. Repici A, Badalamenti M, Maselli R, Correale L, Radaelli F, Rondonotti E, et al. Efficacy of Real-Time Computer-Aided Detection of Colorectal Neoplasia in a Randomized Trial. Gastroenterology. 2020;159(2):512-20.e7.

20. Hassan C, Spadaccini M, Iannone A, Maselli R, Jovani M, Chandrasekar VT, et al. Performance of artificial intelligence in colonoscopy for adenoma and polyp detection: a systematic review and meta-analysis. Gastrointestinal Endoscopy. 2021;93(1):77-85.e6.

21. Karpukhin V, Oğuz B, Min S, Wu L, Edunov S, Chen D, et al. Dense Passage Retrieval for Open-Domain Question Answering2020.

22. Banerjee S, Cash BD, Dominitz JA, Baron TH, Anderson MA, Ben-Menachem T, et al. The role of endoscopy in the management of patients with peptic ulcer disease. Gastrointest Endosc. 2010;71(4):663-8.

23. Johnson J, Douze M, Jégou H. Billion-Scale Similarity Search with GPUs. IEEE Transactions on Big Data. 2017;7.

24. Bender E, Gebru T, McMillan-Major A, Shmitchell S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?2021. 610-23 p.

25. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of Hallucination in Natural Language Generation. ACM Comput Surv. 2023;55(12):Article 248.

26. Amann J, Blasimme A, Vayena E, Frey D, Madai VI. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak. 2020;20(1):310.

27. Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ, et al. Foundation models for generalist medical artificial intelligence. Nature. 2023;616(7956):259-65.

28. Nori H, King N, McKinney S, Carignan D, Horvitz E. Capabilities of GPT-4 on Medical Challenge Problems2023.

29. Pang T, Li P, Zhao L. A survey on automatic generation of medical imaging reports based on deep learning. Biomed Eng Online. 2023;22(1):48.

30. Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, et al. Llama 2: Open Foundation and Fine-Tuned Chat Models2023.

31. Izacard G, Grave E. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering2021. 874-80 p.

32. Peng B, Galley M, He P, Cheng H, Xie Y, Hu Y, et al. Check Your Facts and Try Again: Improving Large Language Models with External Knowledge and Automated Feedback2023.

33. Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, et al. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. ACM Trans Inf Syst. 2025;43(2):Article 42.

34. Reimers N, Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. ArXiv. 2019;abs/1908.10084.

35. Muennighoff N, Tazi N, Magne L, Reimers N. MTEB: Massive Text Embedding Benchmark2022.

36. Thakur N, Reimers N, Rücklé A, Srivastava A, Gurevych I. BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models2021.

37. Ram O, Levine Y, Dalmedigos I, Muhlgay D, Shashua A, Leyton-Brown K, et al. In-Context Retrieval-Augmented Language Models2023.

38. Xu P, Ping W, Wu X, McAfee LC, Zhu C, Liu Z, et al. Retrieval meets Long Context Large Language Models. ArXiv. 2023;abs/2310.03025.

39. Chowdhery A, Narang S, Devlin J, Bosma M, Mishra G, Roberts A, et al. PaLM: Scaling Language Modeling with Pathways2022.

40. Wei J, Tay Y, Bommasani R, Raffel C, Zoph B, Borgeaud S, et al. Emergent Abilities of Large Language Models2022.

41. Luo R, Sun L, Xia Y, Qin T, Zhang S, Poon H, et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining. Briefings in Bioinformatics. 2022;23.

42. Chalasani N, Younossi Z, Lavine JE, Charlton M, Cusi K, Rinella M, et al. The diagnosis and management of nonalcoholic fatty liver disease: Practice guidance from the American Association for the Study of Liver Diseases. Hepatology. 2018;67(1):328-57.

43. Marrero JA, Kulik LM, Sirlin CB, Zhu AX, Finn RS, Abecassis MM, et al. Diagnosis, Staging, and Management of Hepatocellular Carcinoma: 2018 Practice Guidance by the American Association for the Study of Liver Diseases. Hepatology. 2018;68(2):723-50.

44. EASL Clinical Practice Guidelines: Management of hepatocellular carcinoma. J Hepatol. 2018;69(1):182-236.

45. Benson AB, Venook AP, Al-Hawary MM, Arain MA, Chen YJ, Ciombor KK, et al. Colon Cancer, Version 2.2021, NCCN Clinical Practice Guidelines in Oncology. J Natl Compr Canc Netw. 2021;19(3):329-59.

46. Lordick F, Carneiro F, Cascinu S, Fleitas T, Haustermans K, Piessen G, et al. Gastric cancer: ESMO Clinical Practice Guideline for diagnosis, treatment and follow-up. Ann Oncol. 2022;33(10):1005-20.

47. Tenner S, Baillie J, DeWitt J, Vege SS. American College of Gastroenterology guideline: management of acute pancreatitis. Am J Gastroenterol. 2013;108(9):1400-15; 16.

48. Gardner TB, Adler DG, Forsmark CE, Sauer BG, Taylor JR, Whitcomb DC. ACG Clinical Guideline: Chronic Pancreatitis. Am J Gastroenterol. 2020;115(3):322-39.

49. Vege SS, Ziring B, Jain R, Moayyedi P. American gastroenterological association institute guideline on the diagnosis and management of asymptomatic neoplastic pancreatic cysts. Gastroenterology. 2015;148(4):819-22; quize12-3.

50. Ucdal M, Bakhshandehpour A, Durak MB, Balaban Y, Kekilli M, Simsek C. Evaluating the Role of Artificial Intelligence in Making Clinical Decisions for Treating Acute Pancreatitis. Journal of Clinical Medicine. 2025;14(12):4347.

51. Feuerstein JD, Isaacs KL, Schneider Y, Siddique SM, Falck-Ytter Y, Singh S. AGA Clinical Practice Guidelines on the Management of Moderate to Severe Ulcerative Colitis. Gastroenterology. 2020;158(5):1450-61.

52. Lichtenstein GR, Loftus EV, Afzali A, Long MD, Barnes EL, Isaacs KL, et al. ACG Clinical Guideline: Management of Crohn's Disease in Adults. Am J Gastroenterol. 2025;120(6):1225-64.

53. Singh S, Ananthakrishnan AN, Nguyen NH, Cohen BL, Velayos FS, Weiss JM, et al. AGA Clinical Practice Guideline on the Role of Biomarkers for the Management of Ulcerative Colitis. Gastroenterology. 2023;164(3):344-72.

54. Beam AL, Kohane IS. Big Data and Machine Learning in Health Care. Jama. 2018;319(13):1317-8.

55. Cabitza F, Rasoini R, Gensini GF. Unintended Consequences of Machine Learning in Medicine. Jama. 2017;318(6):517-8.

56. Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. N Engl J Med. 2019;380(14):1347-58.

57. Sendak MP, Gao M, Brajer N, Balu S. Presenting machine learning model information to clinical end users with model facts labels. npj Digital Medicine. 2020;3(1):41.

58. Liu Y, Chen PC, Krause J, Peng L. How to Read Articles That Use Machine Learning: Users' Guides to the Medical Literature. Jama. 2019;322(18):1806-16.

59. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 2019;17(1):195.

60. Khattab O, Zaharia M. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT2020. 39-48 p.

61. Yao S, Yu D, Zhao J, Shafran I, Griffiths TL, Cao Y, et al. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. ArXiv. 2023;abs/2305.10601.

62. Schaefer J, Lehne M, Schepers J, Prasser F, Thun S. The use of machine learning in rare diseases: a scoping review. Orphanet Journal of Rare Diseases. 2020;15(1):145.

63. Haendel M, Vasilevsky N, Unni D, Bologa C, Harris N, Rehm H, et al. How many rare diseases are there? Nat Rev Drug Discov. 2020;19(2):77-8.

64. Chung J, Gulcehre C, Cho K, Yere Y. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. 2014.

65. Nguyen GC, Kaplan GG, Harris ML, Brant SR. A National Survey of the Prevalence and Impact of Clostridium difficile Infection Among Hospitalized Inflammatory Bowel Disease Patients. Official journal of the American College of Gastroenterology | ACG. 2008;103(6):1443-50.

66. Pingua B, Sahoo A, Kandpal M, Murmu D, Rautaray J, Barik RK, et al. Medical LLMs: Fine-Tuning vs. Retrieval-Augmented Generation. Bioengineering (Basel). 2025;12(7).

67. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, et al. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. ACM Trans Comput Healthcare. 2021;3(1):Article 2.

68. Umapathi L, Pal A, Sankarasubbu M. Med-HALT: Medical Domain Hallucination Test for Large Language Models2023.

69. Lee P, Bubeck S, Petro J. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. N Engl J Med. 2023;388(13):1233-9.

70. Bates DW, Cullen DJ, Laird N, Petersen LA, Small SD, Servi D, et al. Incidence of adverse drug events and potential adverse drug events. Implications for prevention. ADE Prevention Study Group. Jama. 1995;274(1):29-34.

71. Jiang Z, Xu F, Gao L, Sun Z, Liu Q, Dwivedi-Yu J, et al. Active Retrieval Augmented Generation2023.

72. Liu N, Zhang T, Liang P. Evaluating Verifiability in Generative Search Engines2023. 7001-25 p.

73. Bakken S. AI in health: keeping the human in the loop. J Am Med Inform Assoc. 2023;30(7):1225-6.

74. Shneiderman B. Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. International Journal of Human-Computer Interaction. 2020;36:1-10.

75. Yu K-H, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nature Biomedical Engineering. 2018;2(10):719-31.

76. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright CL, Mishkin P, et al. Training language models to follow instructions with human feedback. Proceedings of the 36th International Conference on Neural Information Processing Systems; New Orleans, LA, USA: Curran Associates Inc.; 2022. p. Article 2011.

77. Perez E, Ringer S, Lukošiūtė K, Nguyen K, Chen E, Heiner S, et al. Discovering Language Model Behaviors with Model-Written Evaluations2022.

78. Kadavath S, Conerly T, Askell A, Henighan T, Drain D, Perez E, et al. Language Models (Mostly) Know What They Know2022.

79. Malkov Y, Yashunin D. Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2016;PP.

80. Jégou H, Douze M, Schmid C. Product Quantization for Nearest Neighbor Search. IEEE transactions on pattern analysis and machine intelligence. 2011;33:117-28.

81. McHugh M. Interrater reliability: the kappa statistic. Biochemia Medica. 2012:276-82.

82. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159-74.

83. Fleiss J, Levin B, Paik M. Statistical Methods for Rates and Proportions, Third Edition. 2004. p. 187-233.

84. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med. 2016;15(2):155-63.

85. Street RL, Jr., Makoul G, Arora NK, Epstein RM. How does communication heal? Pathways linking clinician-patient communication to health outcomes. Patient Educ Couns. 2009;74(3):295-301.

86. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull. 1979;86(2):420-8.

87. De Vet H, Terwee C, Knol D, Bouter L. When to use agreement versus reliability measures. Journal of clinical epidemiology. 2006;59:1033-9.

88. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. Nat Med. 2019;25(1):30-6.

89. Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, et al. Do no harm: a roadmap for responsible machine learning for health care. Nat Med. 2019;25(9):1337-40.

90. Zhu S, Gilbert M, Chetty I, Siddiqui F. The 2021 landscape of FDA-approved artificial intelligence/machine learning-enabled medical devices: An analysis of the characteristics and intended use. International journal of medical informatics. 2022;165:104828.

91. Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. Nature Machine Intelligence. 2019;1(9):389-99.

92. Parikh RB, Teeple S, Navathe AS. Addressing Bias in Artificial Intelligence in Health Care. JAMA. 2019;322(24):2377-8.

93. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Bmj. 2015;350:g7594.

94. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med. 2015;162(1):W1-73.

95. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med. 2022;28(5):924-33.

96. Sounderajah V, Ashrafian H, Aggarwal R, De Fauw J, Denniston AK, Greaves F, et al. Developing specific reporting guidelines for diagnostic accuracy studies

assessing AI interventions: The STARD-AI Steering Group. Nat Med. 2020;26(6):807-8.

97. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Lancet Digit Health. 2020;2(10):e537-e48.

98. Cruz Rivera S, Liu X, Chan AW, Denniston AK, Calvert MJ. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Nat Med. 2020;26(9):1351-63.

99. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. Bmj. 2020;369:m1328.

100. van de Sande D, van Genderen ME, Huiskens J, Gommers D, van Bommel J. Moving from bytes to bedside: a systematic review on the use of artificial intelligence in the intensive care unit. Intensive Care Med. 2021;47(7):750-60.

101. Sambasivan N, Kapania S, Highfill H, Akrong D, Paritosh P, Aroyo L. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI2021. 1-15 p.

102. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med. 2018;1:18.

103. National Academy of M, The Learning Health System S. In: Whicher D, Ahmed M, Israni ST, Matheny M, editors. Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril. Washington (DC): National Academies Press (US) Copyright 2022 by the National Academy of Sciences. All rights reserved.; 2023.

104. Chen RJ, Lu MY, Chen TY, Williamson DFK, Mahmood F. Synthetic data in machine learning for medicine and healthcare. Nat Biomed Eng. 2021;5(6):493-7.

105. Alsentzer E, Murphy J, Boag W, Weng W-H, Jin D, Naumann T, et al. Publicly Available Clinical BERT Embeddings2019.

106. Steinberg E, Jung K, Fries JA, Corbin CK, Pfohl SR, Shah NH. Language models are an effective representation learning technique for electronic health record data. J Biomed Inform. 2021;113:103637.

107. Huang K, Li J, Ranganath R. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission2019.

108. van Aken B, Papaioannou M, Mayrdorfer M, Gers F, Löser A. Clinical Outcome Prediction from Admission Notes using Self-Supervised Knowledge Integration2021.

109. Cai C, Winter S, Steiner D, Wilcox L, Terry M. "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. Proceedings of the ACM on Human-Computer Interaction. 2019;3:1-24.

110. Rajpurkar P, Irvin J, Ball RL, Zhu K, Yang B, Mehta H, et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLoS Med. 2018;15(11):e1002686.

111. Tschandl P, Rinner C, Apalla Z, Argenziano G, Codella N, Halpern A, et al. Human–computer collaboration for skin cancer recognition. Nature Medicine. 2020;26:1229-34.

112. Ehteshami Bejnordi B, Veta M, Diest P, Ginneken B, Karssemeijer N, Litjens G, et al. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. JAMA. 2017;318:2199-210.

113. Abràmoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. npj Digital Medicine. 2018;1(1):39.

114. van der Velden BHM, Kuijf HJ, Gilhuijs KGA, Viergever MA. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. Med Image Anal. 2022;79:102470.

115. Zhang K, Liu X, Shen J, Li Z, Sang Y, Wu X, et al. Clinically Applicable AI System for Accurate Diagnosis, Quantitative Measurements, and Prognosis of COVID-19 Pneumonia Using Computed Tomography. Cell. 2020;181(6):1423-33.e11.

116. Parisi GI, Kemker R, Part JL, Kanan C, Wermter S. Continual lifelong learning with neural networks: A review. Neural Netw. 2019;113:54-71.

117. Howard J, Ruder S. Universal Language Model Fine-tuning for Text Classification2018. 328-39 p.

118. Hu E, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al. LoRA: Low-Rank Adaptation of Large Language Models2021.

119. Gilpin L, Bau D, Yuan B, Bajwa A, Specter M, Kagal L. Explaining Explanations: An Overview of Interpretability of Machine Learning2018. 80-9 p.

120. Ribeiro MT, Singh S, Guestrin C. "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; San Francisco, California, USA: Association for Computing Machinery; 2016. p. 1135–44.

121. Chen IY, Szolovits P, Ghassemi M. Can AI Help Reduce Disparities in General Medical and Mental Health Care? AMA J Ethics. 2019;21(2):E167-79.

122. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. JAMA Intern Med. 2018;178(11):1544-7.

123. Eaneff S, Obermeyer Z, Butte AJ. The Case for Algorithmic Stewardship for Artificial Intelligence and Machine Learning Technologies. Jama. 2020;324(14):1397-8.

124. Maddox TM, Rumsfeld JS, Payne PRO. Questions for Artificial Intelligence in Health Care. Jama. 2019;321(1):31-2.

125. Sanders GD, Neumann PJ, Basu A, Brock DW, Feeny D, Krahn M, et al. Recommendations for Conduct, Methodological Practices, and Reporting of Cost-effectiveness Analyses: Second Panel on Cost-Effectiveness in Health and Medicine. Jama. 2016;316(10):1093-103.

126. Choudhury A, Asan O. Role of Artificial Intelligence in Patient Safety Outcomes: Systematic Literature Review. JMIR Med Inform. 2020;8(7):e18599.

127. Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. J Am Med Inform Assoc. 2012;19(1):121-7.

128. Lyell D, Magrabi F, Raban MZ, Pont LG, Baysari MT, Day RO, et al. Automation bias in electronic prescribing. BMC Med Inform Decis Mak. 2017;17(1):28.

129. Hoff K, Bashir M. Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. Human Factors The Journal of the Human Factors and Ergonomics Society. 2015;57:407-34.

130. Price WN, 2nd, Gerke S, Cohen IG. Potential Liability for Physicians Using Artificial Intelligence. Jama. 2019;322(18):1765-6.

131. Froomkin A, Kerr I, Pineau J. When AIs Outperform Doctors: The Dangers of a Tort-Induced Over-Reliance on Machine Learning and What (Not) to Do About it. SSRN Electronic Journal. 2018.

132. Sullivan HR, Schweikart SJ. Are Current Tort Liability Doctrines Adequate for Addressing Injury Caused by AI? AMA J Ethics. 2019;21(2):E160-6.

133. Asan O, Bayrak AE, Choudhury A. Artificial Intelligence and Human Trust in Healthcare: Focus on Clinicians. J Med Internet Res. 2020;22(6):e15154.

134. Longoni C, Morewedge C. Resistance to medical artificial intelligence is an attribute in a compensatory decision process: response to Pezzo and Beckstead (2020). Judgment and Decision Making. 2020;15:446-8.

135. Richardson JP, Smith C, Curtis S, Watson S, Zhu X, Barry B, et al. Patient apprehensions about the use of artificial intelligence in healthcare. npj Digital Medicine. 2021;4(1):140.

136. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Health. 2019;1(6):e271-e97.

137. Ting DSW, Pasquale LR, Peng L, Campbell JP, Lee AY, Raman R, et al. Artificial intelligence and deep learning in ophthalmology. Br J Ophthalmol. 2019;103(2):167-75.

138. Mandel JC, Kreda DA, Mandl KD, Kohane IS, Ramoni RB. SMART on FHIR: a standards-based, interoperable apps platform for electronic health records. J Am Med Inform Assoc. 2016;23(5):899-908.

139. Rieke N, Hancox J, Li W, Milletarì F, Roth HR, Albarqouni S, et al. The future of digital health with federated learning. NPJ Digit Med. 2020;3:119.

140. Sheller MJ, Edwards B, Reina GA, Martin J, Pati S, Kotrotsou A, et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. Scientific Reports. 2020;10(1):12598.

141. Kaissis GA, Makowski MR, Rückert D, Braren RF. Secure, privacy-preserving and federated machine learning in medical imaging. Nature Machine Intelligence. 2020;2(6):305-11.

142. Jiang X, Osl M, Kim J, Ohno-Machado L. Calibrating predictive model estimates to support personalized medicine. J Am Med Inform Assoc. 2012;19(2):263-74.

143. Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. Proceedings of the 22nd international conference on Machine learning; Bonn, Germany: Association for Computing Machinery; 2005. p. 625–32.

144. Kompa B, Snoek J, Beam AL. Second opinion needed: communicating uncertainty in medical machine learning. NPJ Digit Med. 2021;4(1):4.

145. Reynolds L, McDonell K. Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm2021.

146. Liu P, Yuan W, Fu J, Jiang Z, Hayashi H, Neubig G. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. ACM Computing Surveys. 2022;55.

147. Kojima T, Gu S, Reid M, Matsuo Y, Iwasawa Y. Large Language Models are Zero-Shot Reasoners2022.

148. Sendak MP, Elish MC, Gao M, Futoma JD, Ratliff W, Nichols M, et al. "The human body is a black box": supporting clinical decision-making with deep learning. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 2019.

149. Baylor E, Beede E, Hersch F, Iurchenko A, Ruamviboonsuk P, Vardoulakis L, et al. A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy2020.

150. Coiera E. The Last Mile: Where Artificial Intelligence Meets Reality. J Med Internet Res. 2019;21(11):e16323.

151. Paranjape K, Schinkel M, Nannan Panday R, Car J, Nanayakkara P. Introducing Artificial Intelligence Training in Medical Education. JMIR Med Educ. 2019;5(2):e16048.

152. Kolachalama VB, Garg PS. Machine learning and medical education. NPJ Digit Med. 2018;1:54.

153. Emanuel EJ, Wachter RM. Artificial Intelligence in Health Care: Will the Value Match the Hype? Jama. 2019;321(23):2281-2.

154. Verghese A, Shah NH, Harrington RA. What This Computer Needs Is a Physician: Humanism and Artificial Intelligence. JAMA. 2018;319(1):19-20.

155. Ancker JS, Edwards A, Nosal S, Hauser D, Mauer E, Kaushal R. Effects of workload, work complexity, and repeated alerts on alert fatigue in a clinical decision support system. BMC Med Inform Decis Mak. 2017;17(1):36.

156. van der Sijs H, Aarts J, Vulto A, Berg M. Overriding of drug safety alerts in computerized physician order entry. J Am Med Inform Assoc. 2006;13(2):138-47.

157. Roshanov PS, Fernandes N, Wilczynski JM, Hemens BJ, You JJ, Handler SM, et al. Features of effective computerised clinical decision support systems: meta-regression of 162 randomised trials. Bmj. 2013;346:f657.

158. Klaise J, Looveren A, Cox C, Vacanti G, Coca A. Monitoring and explainability of models in production2020.

159. Feng J, Phillips RV, Malenica I, Bishara A, Hubbard AE, Celi LA, et al. Clinical artificial intelligence quality improvement: towards continual monitoring and updating of AI algorithms in healthcare. NPJ Digit Med. 2022;5(1):66.

160. Saria S, Subbaswamy A. Tutorial: Safe and Reliable Machine Learning2019.

161. Ozer A, Yakupogullari Y, Beytur A, Beytur L, Koroglu M, Salman F, et al. Risk factors of hepatitis B virus infection in Turkey: A population-based, case-control study: Risk Factors for HBV Infection. Hepat Mon. 2011;11(4):263-8.

162. Smits TC, Weru L, Gehlenborg N, L'Yi S. Ten simple rules for making biomedical data resources accessible. PLoS Comput Biol. 2025;21(11):e1013657.

163. Beam AL, Manrai AK, Ghassemi M. Challenges to the Reproducibility of Machine Learning Models in Health Care. Jama. 2020;323(4):305-6.

164. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring Fairness in Machine Learning to Advance Health Equity. Ann Intern Med. 2018;169(12):866-72.

165. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JP. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J Am Med Inform Assoc. 2017;24(1):198-208.

166. Chen JH, Asch SM. Machine Learning and Prediction in Medicine — Beyond the Peak of Inflated Expectations. New England Journal of Medicine. 2017;376(26):2507-9.

167. Ghassemi M, Naumann T, Schulam P, Beam AL, Chen IY, Ranganath R. A Review of Challenges and Opportunities in Machine Learning for Health. AMIA Jt Summits Transl Sci Proc. 2020;2020:191-200.

168. Topol EJ. Welcoming new guidelines for AI clinical research. Nat Med. 2020;26(9):1318-20.

169. Cabitza F, Campagner A, Balsano C. Bridging the "last mile" gap between AI implementation and operation: "data awareness" that matters. Ann Transl Med. 2020;8(7):501.

170. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc J. 2019;6(2):94-8.

171. Panch T, Szolovits P, Atun R. Artificial intelligence, machine learning and health systems. J Glob Health. 2018;8(2):020303.

172. Stead WW. Clinical Implications and Challenges of Artificial Intelligence and Deep Learning. Jama. 2018;320(11):1107-8.

173. Zhou H, Yu X, Alhaskawi A, Dong Y, Wang Z, Jin Q, et al. A deep learning approach for medical waste classification. Sci Rep. 2022;12(1):2159.

174. Gao J, Lanchantin J, Qi Y. Black-Box Generation of Adversarial Text Sequences to Evade Deep Learning Classifiers2018. 50-6 p.

175. Rs R, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. International Journal of Computer Vision. 2020;128.

176. Kather JN, Pearson AT, Halama N, Jäger D, Krause J, Loosen SH, et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. Nat Med. 2019;25(7):1054-6.

177. Iizuka O, Kanavati F, Kato K, Rambeau M, Arihiro K, Tsuneki M. Deep Learning Models for Histopathological Classification of Gastric and Colonic Epithelial Tumours. Sci Rep. 2020;10(1):1504.

178. Korbar B, Olofson AM, Miraflor AP, Nicka CM, Suriawinata MA, Torresani L, et al. Deep Learning for Classification of Colorectal Polyps on Whole-slide Images. J Pathol Inform. 2017;8:30.

179. Urban G, Tripathi P, Alkayali T, Mittal M, Jalali F, Karnes W, et al. Deep Learning Localizes and Identifies Polyps in Real Time With 96% Accuracy in Screening Colonoscopy. Gastroenterology. 2018;155(4):1069-78.e8.

180. Byrne MF, Chapados N, Soudan F, Oertel C, Linares Pérez M, Kelly R, et al. Real-time differentiation of adenomatous and hyperplastic diminutive colorectal polyps during analysis of unaltered videos of standard colonoscopy using a deep learning model. Gut. 2019;68(1):94-100.

181. Wang P, Xiao X, Glissen Brown JR, Berzin TM, Tu M, Xiong F, et al. Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy. Nat Biomed Eng. 2018;2(10):741-8.

182. Misawa M, Kudo SE, Mori Y, Hotta K, Ohtsuka K, Matsuda T, et al. Development of a computer-aided detection system for colonoscopy and a publicly accessible large colonoscopy video database (with video). Gastrointest Endosc. 2021;93(4):960-7.e3.

183. Vinsard DG, Mori Y, Misawa M, Kudo SE, Rastogi A, Bagci U, et al. Quality assurance of computer-aided detection and diagnosis in colonoscopy. Gastrointest Endosc. 2019;90(1):55-63.

184. Rex DK, Boland CR, Dominitz JA, Giardiello FM, Johnson DA, Kaltenbach T, et al. Colorectal Cancer Screening: Recommendations for Physicians and Patients from the U.S. Multi-Society Task Force on Colorectal Cancer. Am J Gastroenterol. 2017;112(7):1016-30.

185. Brenner H, Stock C, Hoffmeister M. Effect of screening sigmoidoscopy and screening colonoscopy on colorectal cancer incidence and mortality: systematic review and meta-analysis of randomised controlled trials and observational studies. Bmj. 2014;348:g2467.

186. Robertson DJ, Lee JK, Boland CR, Dominitz JA, Giardiello FM, Johnson DA, et al. Recommendations on Fecal Immunochemical Testing to Screen for Colorectal Neoplasia: A Consensus Statement by the US Multi-Society Task Force on Colorectal Cancer. Gastroenterology. 2017;152(5):1217-37.e3.

187. Shaukat A, Kahi CJ, Burke CA, Rabeneck L, Sauer BG, Rex DK. ACG Clinical Guidelines: Colorectal Cancer Screening 2021. Am J Gastroenterol. 2021;116(3):458-79.

188. Davidson KW, Barry MJ, Mangione CM, Cabana M, Caughey AB, Davis EM, et al. Screening for Colorectal Cancer: US Preventive Services Task Force Recommendation Statement. Jama. 2021;325(19):1965-77.

189. Steyerberg E. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating2009.

190. Davis SE, Lasko TA, Chen G, Siew ED, Matheny ME. Calibration drift in regression and machine learning models for acute kidney injury. J Am Med Inform Assoc. 2017;24(6):1052-61.

191. Finlayson SG, Subbaswamy A, Singh K, Bowers J, Kupke A, Zittrain J, et al. The Clinician and Dataset Shift in Artificial Intelligence. N Engl J Med. 2021;385(3):283-6.

192. Stevens LM, Mortazavi BJ, Deo RC, Curtis L, Kao DP. Recommendations for Reporting Machine Learning Analyses in Clinical Research. Circ Cardiovasc Qual Outcomes. 2020;13(10):e006556.

193. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, et al. Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View. J Med Internet Res. 2016;18(12):e323.

194. Riley RD, Ensor J, Snell KIE, Harrell FE, Jr., Martin GP, Reitsma JB, et al. Calculating the sample size required for developing a clinical prediction model. Bmj. 2020;368:m441.

195. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. Eur Heart J. 2014;35(29):1925-31.

196. Moons KG, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart. 2012;98(9):691-8.

197. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17(1):230.

198. Raine T, Bonovas S, Burisch J, Kucharzik T, Adamina M, Annese V, et al. ECCO Guidelines on Therapeutics in Ulcerative Colitis: Medical Treatment. J Crohns Colitis. 2022;16(1):2-17.

199. Olivera P, Danese S, Peyrin-Biroulet L. Next generation of small molecules in inflammatory bowel disease. Gut. 2017;66(2):199-209.

200. Mohammed Vashist N, Samaan M, Mosli MH, Parker CE, MacDonald JK, Nelson SA, et al. Endoscopic scoring indices for evaluation of disease activity in ulcerative colitis. Cochrane Database Syst Rev. 2018;1(1):Cd011450.

201. EASL Clinical Practice Guidelines on non-invasive tests for evaluation of liver disease severity and prognosis - 2021 update. J Hepatol. 2021;75(3):659-89.

202. Younossi ZM, Loomba R, Anstee QM, Rinella ME, Bugianesi E, Marchesini G, et al. Diagnostic modalities for nonalcoholic fatty liver disease, nonalcoholic steatohepatitis, and associated fibrosis. Hepatology. 2018;68(1):349-60.

203. Castera L, Friedrich-Rust M, Loomba R. Noninvasive Assessment of Liver Disease in Patients With Nonalcoholic Fatty Liver Disease. Gastroenterology. 2019;156(5):1264-81.e4.

204. Hill ID, Dirks MH, Liptak GS, Colletti RB, Fasano A, Guandalini S, et al. Guideline for the diagnosis and treatment of celiac disease in children: recommendations of the North American Society for Pediatric Gastroenterology, Hepatology and Nutrition. J Pediatr Gastroenterol Nutr. 2005;40(1):1-19.

205. Wine E, Aloi M, Van Biervliet S, Bronsky J, di Carpi JM, Gasparetto M, et al. Management of paediatric ulcerative colitis, part 1: Ambulatory care-An updated evidence-based consensus guideline from the European Society of Paediatric Gastroenterology, Hepatology and Nutrition and the European Crohn's and Colitis Organisation. J Pediatr Gastroenterol Nutr. 2025;81(3):765-815.

206. Koletzko S, Niggemann B, Arato A, Dias JA, Heuschkel R, Husby S, et al. Diagnostic approach and management of cow's-milk protein allergy in infants and children: ESPGHAN GI Committee practical guidelines. J Pediatr Gastroenterol Nutr. 2012;55(2):221-9.

207. Vogel A, Cervantes A, Chau I, Daniele B, Llovet JM, Meyer T, et al. Hepatocellular carcinoma: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol. 2018;29(Suppl 4):iv238-iv55.

208. Drossman DA, Hasler WL. Rome IV-Functional GI Disorders: Disorders of Gut-Brain Interaction. Gastroenterology. 2016;150(6):1257-61.

209. Mearin F, Lacy BE, Chang L, Chey WD, Lembo AJ, Simren M, et al. Bowel Disorders. Gastroenterology. 2016.

210. Ford AC, Sperber AD, Corsetti M, Camilleri M. Irritable bowel syndrome. Lancet. 2020;396(10263):1675-88.

211. Strate LL, Gralnek IM. ACG Clinical Guideline: Management of Patients With Acute Lower Gastrointestinal Bleeding. Am J Gastroenterol. 2016;111(4):459-74.

212. Barkun AN, Almadi M, Kuipers EJ, Laine L, Sung J, Tse F, et al. Management of Nonvariceal Upper Gastrointestinal Bleeding: Guideline Recommendations

From the International Consensus Group. Ann Intern Med. 2019;171(11):805-22.

213. Banks PA, Bollen TL, Dervenis C, Gooszen HG, Johnson CD, Sarr MG, et al. Classification of acute pancreatitis--2012: revision of the Atlanta classification and definitions by international consensus. Gut. 2013;62(1):102-11.

214. Cederholm T, Jensen GL, Correia M, Gonzalez MC, Fukushima R, Higashiguchi T, et al. GLIM criteria for the diagnosis of malnutrition - A consensus report from the global clinical nutrition community. Clin Nutr. 2019;38(1):1-9.

215. Bischoff SC, Austin P, Boeykens K, Chourdakis M, Cuerda C, Jonkers-Schuitema C, et al. ESPEN guideline on home enteral nutrition. Clin Nutr. 2020;39(1):5-22.

216. Pironi L, Arends J, Bozzetti F, Cuerda C, Gillanders L, Jeppesen PB, et al. ESPEN guidelines on chronic intestinal failure in adults. Clin Nutr. 2016;35(2):247-307.

217. Chandrasekhara V, Khashab MA, Muthusamy VR, Acosta RD, Agrawal D, Bruining DH, et al. Adverse events associated with ERCP. Gastrointest Endosc. 2017;85(1):32-47.

218. Adler DG, Lieb JG, 2nd, Cohen J, Pike IM, Park WG, Rizk MK, et al. Quality indicators for ERCP. Gastrointest Endosc. 2015;81(1):54-66.

219. Biggins SW, Angeli P, Garcia-Tsao G, Ginès P, Ling SC, Nadim MK, et al. Diagnosis, Evaluation, and Management of Ascites, Spontaneous Bacterial Peritonitis and Hepatorenal Syndrome: 2021 Practice Guidance by the American Association for the Study of Liver Diseases. Hepatology. 2021;74(2):1014-48.

220. Kamath PS, Kim WR. The model for end-stage liver disease (MELD). Hepatology. 2007;45(3):797-805.

221. Yadlapati R, Kahrilas PJ, Fox MR, Bredenoord AJ, Prakash Gyawali C, Roman S, et al. Esophageal motility disorders on high-resolution manometry: Chicago classification version 4.0(©). Neurogastroenterol Motil. 2021;33(1):e14058.

222. Camilleri M, Parkman HP, Shafi MA, Abell TL, Gerson L. Clinical guideline: management of gastroparesis. Am J Gastroenterol. 2013;108(1):18-37; quiz 8.

223. Sculley D, Holt G, Golovin D, Davydov E, Phillips T, Ebner D, et al. Hidden Technical Debt in Machine Learning Systems. NIPS. 2015:2494-502.

224. Wagstaff K. Machine Learning that Matters. Proceedings of 29th International Conference on Machine Learning. 2012;1.

225. Howard J, Gugger S. fastai: A Layered API for Deep Learning. Inf. 2020;11:108.

226. Nundy S, Montgomery T, Wachter RM. Promoting Trust Between Patients and Physicians in the Era of Artificial Intelligence. Jama. 2019;322(6):497-8.

227. Shen J, Zhang CJP, Jiang B, Chen J, Song J, Liu Z, et al. Artificial Intelligence Versus Clinicians in Disease Diagnosis: Systematic Review. JMIR Med Inform. 2019;7(3):e10010.

228. Allyn J, Allou N, Augustin P, Philip I, Martinet O, Belghiti M, et al. A Comparison of a Machine Learning Model with EuroSCORE II in Predicting Mortality after Elective Cardiac Surgery: A Decision Curve Analysis. PLoS One. 2017;12(1):e0169772.

229. Wu E, Wu K, Daneshjou R, Ouyang D, Ho DE, Zou J. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. Nat Med. 2021;27(4):582-4.

230. Meskó B, Hetényi G, Győrffy Z. Will artificial intelligence solve the human resource crisis in healthcare? BMC Health Serv Res. 2018;18(1):545.

231. Lin SY, Mahoney MR, Sinsky CA. Ten Ways Artificial Intelligence Will Transform Primary Care. J Gen Intern Med. 2019;34(8):1626-30.

232. Coravos A, Khozin S, Mandl KD. Developing and adopting safe and effective digital biomarkers to improve patient outcomes. NPJ Digit Med. 2019;2(1).

233. Lehman CD, Wellman RD, Buist DS, Kerlikowske K, Tosteson AN, Miglioretti DL. Diagnostic Accuracy of Digital Screening Mammography With and Without Computer-Aided Detection. JAMA Intern Med. 2015;175(11):1828-37.

234. Coiera E, Ammenwerth E, Georgiou A, Magrabi F. Does health informatics have a replication crisis? J Am Med Inform Assoc. 2018;25(8):963-8.

235. Sittig DF, Wright A, Osheroff JA, Middleton B, Teich JM, Ash JS, et al. Grand challenges in clinical decision support. J Biomed Inform. 2008;41(2):387-92.

236. Parikh RB, Obermeyer Z, Navathe AS. Regulation of predictive analytics in medicine. Science. 2019;363(6429):810-2.

237. Schwalbe N, Wahl B. Artificial intelligence and the future of global health. Lancet. 2020;395(10236):1579-86.

238. Wahl B, Cossy-Gantner A, Germann S, Schwalbe NR. Artificial intelligence (AI) and global health: how can AI contribute to health in resource-poor settings? BMJ Glob Health. 2018;3(4):e000798.

239. Char DS, Abràmoff MD, Feudtner C. Identifying Ethical Considerations for Machine Learning Healthcare Applications. Am J Bioeth. 2020;20(11):7-17.

240. Grote T, Berens P. On the ethics of algorithmic decision-making in healthcare. J Med Ethics. 2020;46(3):205-11.

241. Morley J, Floridi L, Kinsey L, Elhalal A. From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. Sci Eng Ethics. 2020;26(4):2141-68.

242. Han ER, Yeo S, Kim MJ, Lee YH, Park KH, Roh H. Medical education trends for future physicians in the era of advanced technology and artificial intelligence: an integrative review. BMC Med Educ. 2019;19(1):460.

243. Wartman SA, Combs CD. Medical Education Must Move From the Information Age to the Age of Artificial Intelligence. Acad Med. 2018;93(8):1107-9.

244. Robbins T, Kyrou I, Arvanitis TN, Randeva HS, Sankar S, Sutherland S, et al. Education and Training: Topol digital fellowship aspirants: Understanding the motivations, priorities and experiences of the next generation of digital health leaders. Future Healthcare Journal. 2022;9(1):51-6.

245. Obermeyer Z, Emanuel EJ. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. N Engl J Med. 2016;375(13):1216-9.

246. Rajpurkar P, Lungren MP. The Current and Future State of AI Interpretation of Medical Images. N Engl J Med. 2023;388(21):1981-90.

247. Floridi L, Cowls J, King TC, Taddeo M. How to Design AI for Social Good: Seven Essential Factors. Sci Eng Ethics. 2020;26(3):1771-96.

248. Mittelstadt B, Allo P, Taddeo M, Wachter S, Floridi L. The Ethics of Algorithms: Mapping the Debate. Big Data & Society. 2016;In press.

249. Shneiderman B. Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-centered AI Systems. ACM Transactions on Interactive Intelligent Systems. 2020;10:1-31.

250. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health. 2021;3(11):e745-e50.

251. Vollmer S, Mateen BA, Bohner G, Király FJ, Ghani R, Jonsson P, et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. Bmj. 2020;368:l6927.

252. Ouanes K, Farhah N. Effectiveness of Artificial Intelligence (AI) in Clinical Decision Support Systems and Care Delivery. J Med Syst. 2024;48(1):74.

253. Qi Y, Mohamad E, Azlan AA, Zhang C. Utilization of artificial intelligence in clinical practice: A systematic review of China's experiences. Digit Health. 2025;11:20552076251343752.

254. Salas M, Petracek J, Yalamanchili P, Aimer O, Kasthuril D, Dhingra S, et al. The Use of Artificial Intelligence in Pharmacovigilance: A Systematic Review of the Literature. Pharmaceut Med. 2022;36(5):295-306.