ÖZET

Doğan T. Kalıtsal Kanser Sendromlarının Genetik Danışmanlığında Üretken Yapay Zeka Araçlarının Bilgi Sağlayıcılığı Açısından Karşılaştırılması. İstanbul Üniversitesi Sağlık Bilimleri Enstitüsü, Tıbbi Genetik Anabilim Dalı, Genetik Danışma Programı. Yüksek Lisans Tezi. İstanbul. 2026.

Bu çalışmada, üretken yapay zeka araçlarının (ÜYZA), kalıtsal kanser sendromları (KKS) bağlamında genetik danışma (GD) oturumlarına hazırlık sürecinde genetik danışmanlara destekleyici bilgi kaynağı olarak sağlayabileceği katkının karşılaştırmalı olarak değerlendirilmesi amaçlandı. Mart 2026 itibarıyla dünyada en çok kullanılan altı ÜYZA, sırasıyla ChatGPT, Gemini, Claude, Grok, DeepSeek ve Copilot, ücretsiz planlarında sunulan en güncel modelleriyle üç ayrı analizde değerlendirildi. İlk aşamada güncel kılavuzlara dayanan 60 açık uçlu soru, ikinci aşamada PubMed'den seçilmiş case report'lardan uyarlanan 60 vaka öyküsü, standart promptlarla ÜYZA'lara yöneltildi. Yanıtlar doktora seviyesinde iki bağımsız genetik uzmanı tarafından skorlandı. Üçüncü aşamada vaka öykülerinden Doc2Hpo yardımıyla türetilen Human Phenotype Ontology (HPO) ID'leri, PubCaseFinder ve ÜYZA'lara verilerek ilk yanıtta doğru final tanıya ulaşılıp ulaşılamadığı ikili değişken olarak kaydedildi. İstatistiksel analizlerde, açık uçlu soru ve vaka öyküsü yanıtlarına ait birleşik skorlar Friedman test ile karşılaştırıldı, ikili karşılaştırmalarda Wilcoxon signed-rank test kullanıldı ve çoklu karşılaştırmalar Holm correction ile düzeltildi. Değerlendiriciler arası uyum kuadratik ağırlıklı Cohen kappa ile incelendi. HPO temelli tanı analizinde araçlar arasındaki genel fark Cochran's Q test ile, PubCaseFinder ile ÜYZA'lar arasındaki ikili karşılaştırmalar ise exact McNemar test ile değerlendirildi. Açık uçlu soru analizinde Gemini ve DeepSeek en yüksek medyan skorlara ulaştı (her ikisi de 8), Grok ve Claude orta düzeyde (medyan 7), Copilot ve ChatGPT en düşük düzeyde (medyan 6) kaldı. ÜYZA'lar arası fark istatistiksel olarak anlamlıydı (Friedman test, chi-square = 252,59, df = 5, $p < 0,0001$) ve değerlendiriciler arası uyum orta düzeydeydi (kappa = 0,484). Vaka öyküsü analizinde Claude en yüksek skorlarla (medyan 8) belirgin biçimde öne çıktı, Grok ve DeepSeek onu izledi, ChatGPT, Gemini ve Copilot geride kaldı. Bu analizde de ÜYZA'lar arası genel fark istatistiksel olarak anlamlıydı (Friedman test, chi-square = 140,90, df = 5, $p < 0,0001$), ancak değerlendiriciler arası uyum belirgin biçimde düşüktü (kappa = -0,096). Bu bulgu, vaka temelli yanıtların değerlendirilmesinin daha yoruma açık olduğunu düşündürdü. HPO temelli tanı analizinde en yüksek doğruluk PubCaseFinder'da saptandı (%20). Claude %16,7 ile onu izledi. ChatGPT %10 doğruluk oranına ulaşırken, Copilot, DeepSeek, Grok ve Gemini %10'un altında kaldı. Araçlar arası genel fark anlamlı bulundu (Cochran's Q test, Q = 23,17, df = 6, p = 0,0007), ancak çoklu karşılaştırma düzeltmesi sonrası PubCaseFinder ile hiçbir ÜYZA arasındaki ikili fark anlamlı değildi (Holm-adjusted $p > 0,005$). Bulgular, tek bir ÜYZA'nın tüm görevlerde üstün olmadığını, performansın görev türüne, yani bilgi sağlama, klinik vaka yorumlama ve ontoloji temelli tanı tahmini düzeylerine göre belirgin biçimde değiştiğini gösterdi. Çalışmanın sınırlı örneklem büyüklüğü, yalnızca ilk yanıtların değerlendirilmesi, ÜYZA'ların hızla güncellenen doğası nedeniyle bulguların ilgili döneme özgü model versiyonlarına ait olması ve gerçek danışan görüşmelerindeki empati, etik ve otonomi boyutlarının kapsam dışı bırakılması başlıca limitasyonlarıdır. ÜYZA'ların genetik danışmanlara güvenilir/verimli şekilde bilgi sağlayabilme potansiyeli yüksektir ancak bu araçlar bağımsız karar vericiler olarak değil, denetim altında kullanılması gereken destekleyici bilgi kaynakları olarak düşünülmelidir.

Anahtar Kelimeler: Genetik danışmanlık, kalıtsal kanser sendromları, üretken yapay zeka araçları, Human Phenotype Ontology, PubCaseFinder.