Onkolojide Yapay Zeka

YAPAY ZEKâ KLİNİK KARAR VERMEYE HAZIR MI? PROSTAT KANSERİNDE BÜYÜK DİL MODELLERİNİN HEKİM REFERANSINA KARŞI PERFORMANSI

Gökhan Çolak[1], Ümit Bulut[2], İrem Aşık Bal[3], Sercan Özbek[1], Dilara Özgül[1], Ayhan Açlan[4], Muhammed Ali Kaypak[4], Bilgin Demir[1], Merve Turan[1], Esin Oktay[1]
[1]Adnan Menderes Üniversitesi Tıp Fakültesi, Tıbbi Onkoloji Bilim Dalı, Aydın
[2]Ege Üniversitesi Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Ana Bilim Dalı, İzmir
[3]Adnan Menderes Üniversitesi Tıp Fakültesi, İç Hastalıkları Ana Bilim Dalı, Aydın
[4]T.c. Sağlık Bakanlığı Aydın Atatürk Devlet Hastanesi, Tıbbi Onkoloji Kliniği, Aydın

Giriş-Amaç

Amaç: Yapay zekâ ve özellikle büyük dil modelleri (LLM), onkolojide tanı, prognostik değerlendirme ve tedavi planlamasına yönelik karar destek potansiyeli ile giderek daha fazla ilgi görmektedir. Bununla birlikte, gerçek yaşam klinik verileri karşısındaki doğrulukları ve uzman kararlarıyla uyumları henüz yeterince gösterilmemiştir. Bu çalışmada, metastatik hormon duyarlı prostat kanseri (mHSPC) yönetiminde dört LLM'nin (ChatGPT 5.2, Gemini 3, Claude, DeepSeek) klinik kararlarını 3 kişilik tıbbi onkoloji konseyi kararlarıyla karşılaştırarak uyum değerlendirmesi amaçlanmıştır.

Yöntem: Adnan Menderes Üniversitesi'nde takip edilen de novo mHSPC'li 193 hastanın verileri anonimleştirildi ve tedavi kararı için standart formlara aktarıldı. Modeller ve konseyden; volüm/risk sınıflaması, tedavi amacı (küratif/palyatif) ve tedavi önerisi değerlendirmesi istendi; konsey her hasta için tek ortak karar verdi. Uyum Cohen kappa ve yüzde uyumla analiz edildi; $p<0,01$ anlamlı kabul edildi.

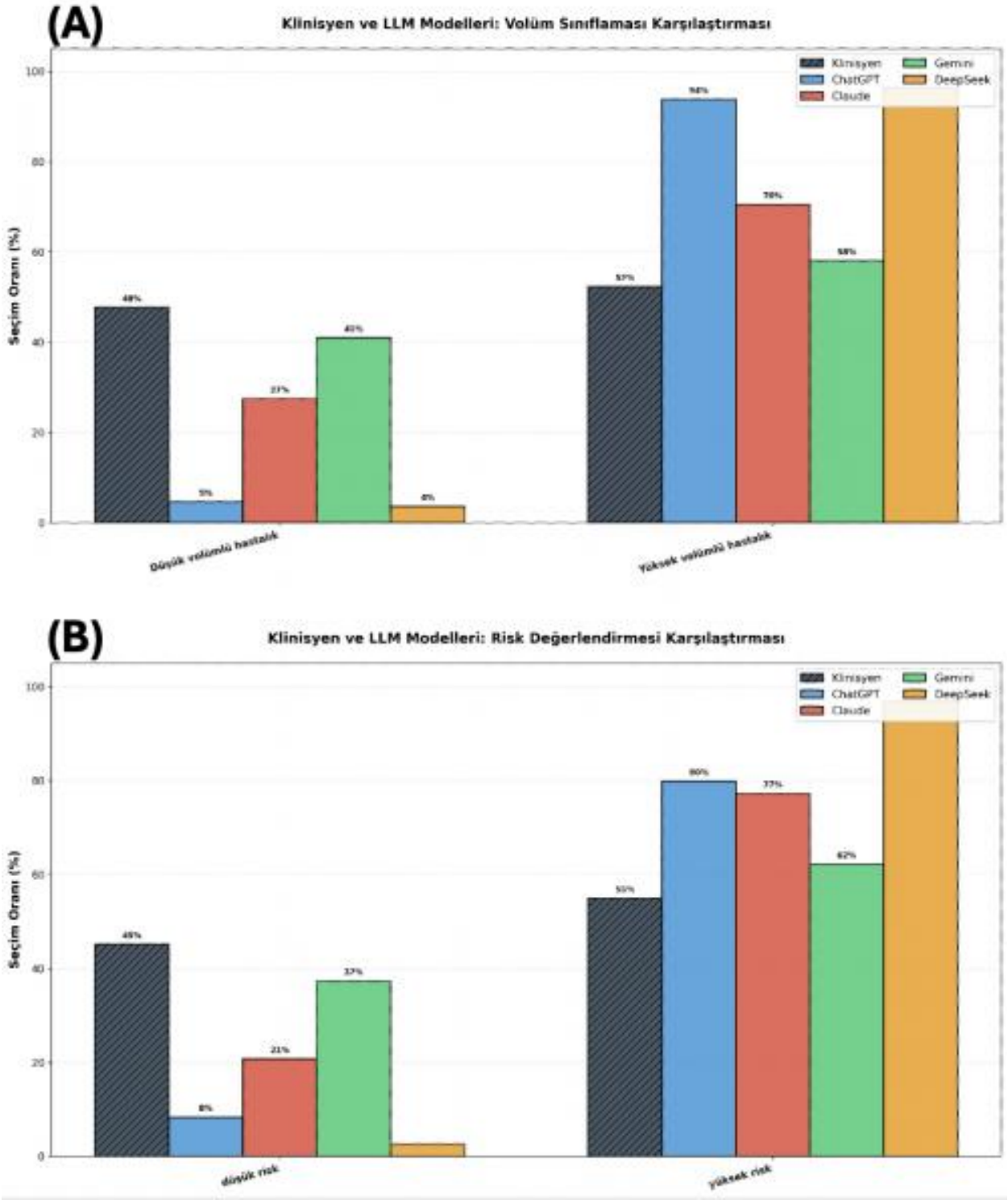
Bulgular: Klinisyen referansına göre volüm değerlendirmesinde (CHAARTED) en yüksek uyum Gemini'de saptandı (κ=0,814; %90,7); Claude orta düzeyde kaldı (κ=0,544; %77,2). ChatGPT (κ=0,096; %55,9) ve DeepSeek (κ=0,058; %54,9) düşük uyum gösterdi ve düşük volüm olgularını çoğunlukla yüksek volüm olarak sınıflandırdı. Risk sınıflamasında (LATITUDE) benzer şekilde Gemini yüksek uyum sağladı (κ=0,768; %88,6), Claude orta düzeydeydi (κ=0,463; %74,1); ChatGPT (κ=0,181; %62,2) ve DeepSeek (κ=0,069; %57,5) zayıf kaldı. Klinisyen 25 hastada küratif hedef belirlemişken, ChatGPT/Gemini/DeepSeek hiç küratif önermezken Claude yalnızca 3 olguda eşleşti. Sistemik tedavi seçiminde en yüksek uyum Claude'da (κ=0,532; %73,6) olup Gemini (κ=0,412; %65,8), ChatGPT (κ=0,302; %59,1) ve DeepSeek (κ=0,150; %58,5) izledi. Tüm modeller ADT tekli tedaviyi belirgin biçimde az tanıyıp ikili/üçlü rejimlere kaydı; DeepSeek ikili, ChatGPT ve Gemini üçlü rejimleri orantısız sıklıkta önerdi. ARPI alt tip eşleşmesi Abirateron'da görece daha iyi, Apalutamid/Darolutamid'de ise yok denecek düzeydeydi.

Sonuç: mHSPC yönetiminde LLM–klinisyen uyumu heterojendi: Gemini volüm/riskte, Claude tedavi seçiminde daha iyi; ChatGPT/DeepSeek yanlış sınıflama ve rejim yoğunlaştırma eğilimindeydi. Bu çalışma; LLM'lerin klinik kararlarda bağımsız kullanım için henüz erken olduğunu, onkolojide ancak uzman denetiminde destekleyici araç olabileceklerini göstermektedir.

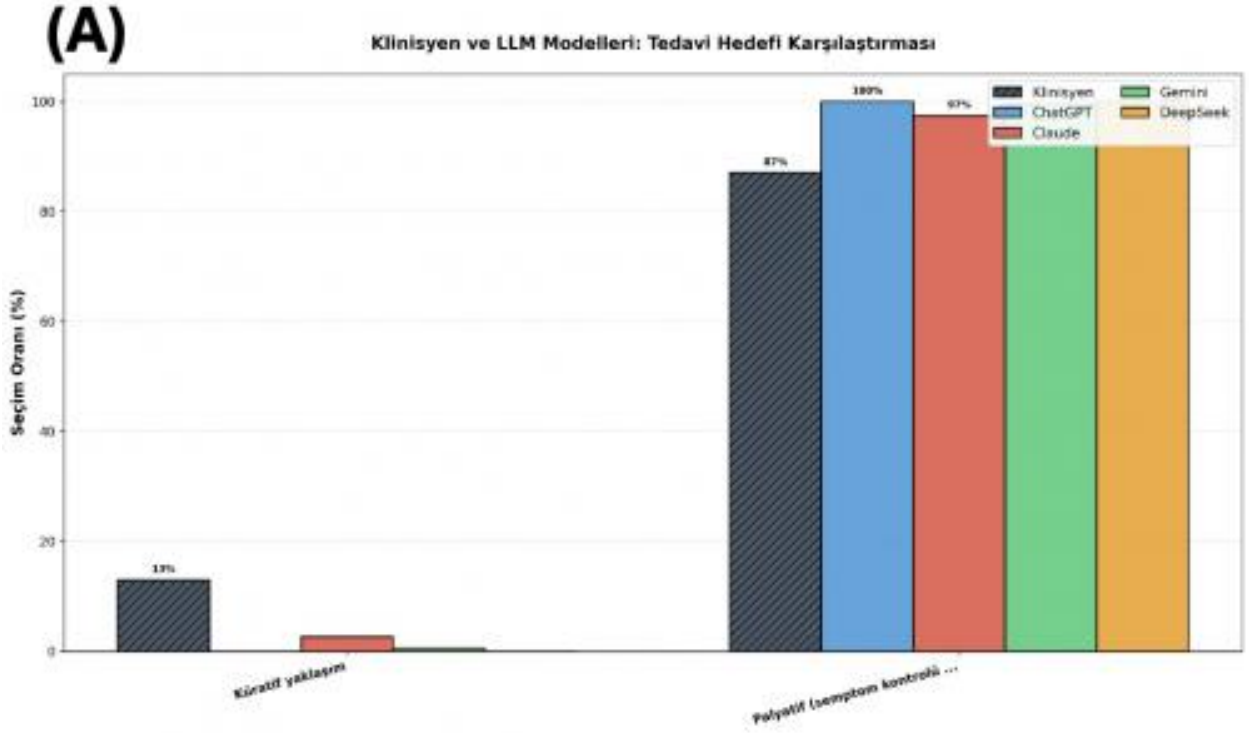
Anahtar Kelimeler : Büyük dil modelleri, Prostat kanseri, Klinik karar uyumu

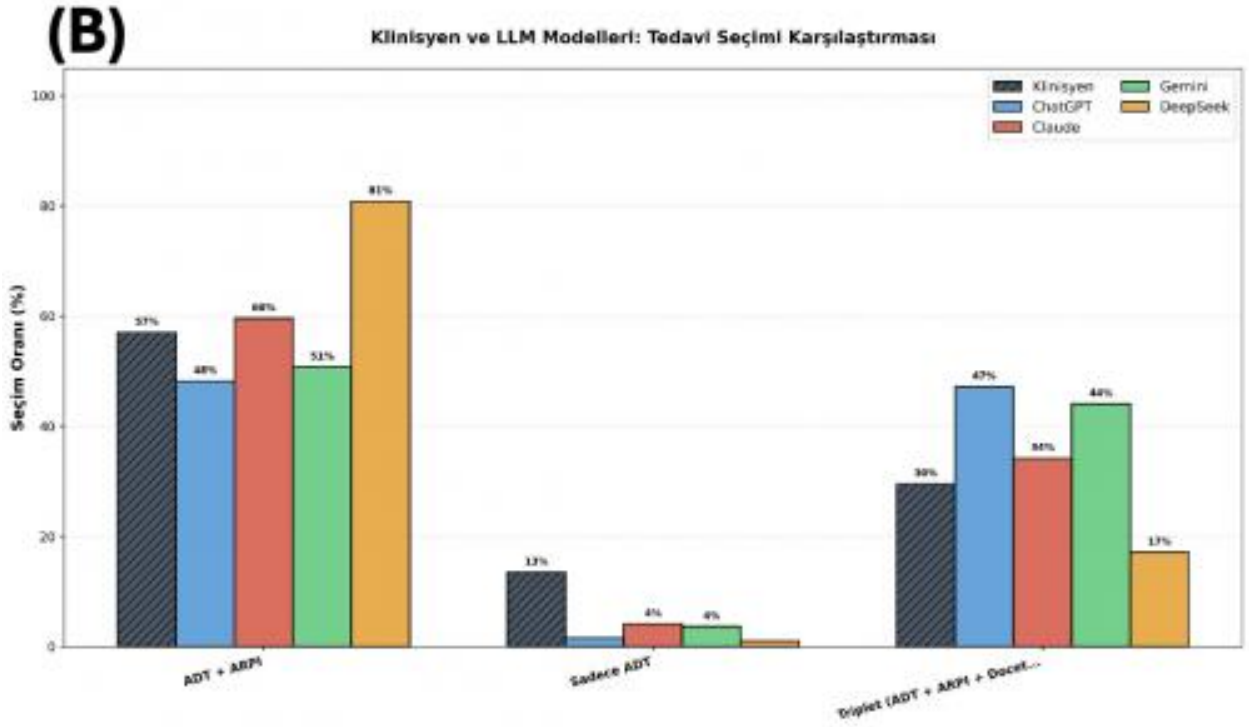
Resimler :

Resim Açıklaması: Şekil 1. Klinisyen ve LLM Modelleri (A) CHAARTED kriterlerine göre volüm değerlendirmesi seçim oranları, (B) LATITUDE kriterlerine göre risk değerlendirmesi seçim oranları.

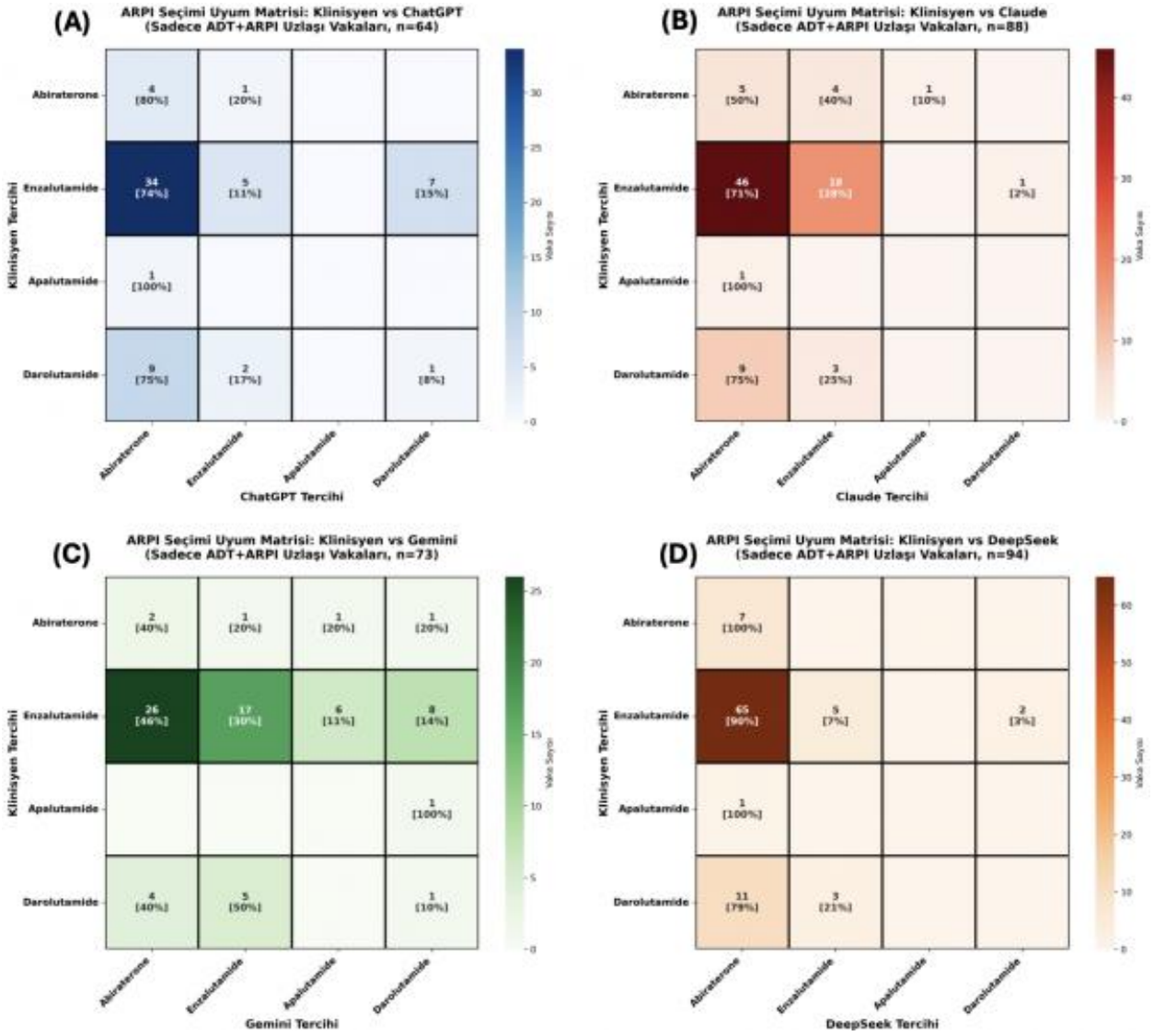


Resim Açıklaması:





Resim Açıklaması:



Tables :

Tablo 1. LLM modellerinin hastalık volüm değerlendirmesinde klinisyen referansına göre performansı

		Klinisyen	Klinisyen	Klinisyen			
		Düşük Volümlü Hastalık	Yüksek Volümlü Hastalık	Belirsiz/ Sınıflandırılamadı	Kappa	p değeri	Yüzde Uyum
		n(%)	n(%)	n(%)			
ChatGPT	Düşük Volümlü Hastalık	9 (9,8)	0(,0)	0(,0)	0,096	0,003	55,9

	Yüksek Volümlü Hastalık	82(89,1)	99(98,0)	0(,0)			
	Belirsiz/ Sınıflandırılamadı	0(,0)	1(1,0)	0(,0)			
Claude	Düşük Volümlü Hastalık	51(55,4)	2(2,0)	0(,0)	0,544	<0,001	77,2
	Yüksek Volümlü Hastalık	38(41,3)	98(97,0)	0(,0)			
	Belirsiz/ Sınıflandırılamadı	3(3,3)	1(1,0)	0(,0)			
Gemini	Düşük Volümlü Hastalık	77(83,7)	2(2,0)	0(,0)	0,814	<0,001	90,7
	Yüksek Volümlü Hastalık	14(15,2)	98(97,0)	0(,0)			
	Belirsiz/ Sınıflandırılamadı	1(1,1)	1(1,0)	0(,0)			
DeepSeek	Düşük Volümlü Hastalık	6(6,5)	1(1,0)	0(,0)	0,058	0,040	54,9
	Yüksek Volümlü Hastalık	86(93,5)	100(99,0)	0(,0)			
	Belirsiz/ Sınıflandırılamadı	0(,0)	0(,0)	0(,0)			

Tablo 2. LLM modellerinin hastalık risk değerlendirmesinde klinisyen referansına göre performansı

		Klinisyen	Klinisyen	Klinisyen			
		Düşük Risk	Yüksek Risk	Belirsiz/Sınıflandırılamadı	Kappa	p değeri	Yüzde Uyum
		n(%)	n(%)	n(%)			
ChatGPT	Düşük Risk	15(17,2)	1(,9)	0(,0)	0,181	<0,001	62,2
	Yüksek Risk	71(81,6)	105(99,1)	0(,0)			
	Belirsiz/Sınıflandırılamadı	1(1,1)	0(,0)	0(,0)			

Claude	Düşük Risk	39(44,8)	1(,9)	0(,0)	0,463	<0,001	74,1
	Yüksek Risk	45(51,7)	104(98,1)	0(,0)			
	Belirsiz/Sınıflandırılamadı	3(3,4)	1(,9)	0(,0)			
Gemini	Düşük Risk	69(79,3)	3(2,8)	0(,0)	0,768	<0,001	88,6
	Yüksek Risk	18(20,7)	102(96,2)	0(,0)			
	Belirsiz/Sınıflandırılamadı	0(,0)	1(,9)	0(,0)			
DeepSeek	Düşük Risk	5(5,7)	0(,0)	0(,0)	0,069	0,007	57,5
	Yüksek Risk	81(93,1)	106(100,0)	0(,0)			
	Belirsiz/Sınıflandırılamadı	1(1,1)	0(,0)	0(,0)			

Tablo 3. LLM modellerinin sistemik tedavi seçiminde klinisyen referansına göre performansı ve uyumu

		Klinisyen	Klinisyen	Klinisyen	Klinisyen	Klinisyen	Klinisyen	Klinisyen			
		ADT	ADT+Docetaksel	ADT+ARPİ	ADT+ARPİ+Docetaksel	ADT+Lokal Tedavi	İzlem	Diğer	Kappa	p değeri	Yüzde Uyum
		n(%)	n(%)	n(%)	n(%)	n(%)	n(%)	n(%)			
Chat GPT	ADT	2(8,0)	0(,0)	1(,9)	0(,0)	0(,0)	0(,0)	0(,0)	0,302	<0,001	59,1
	ADT+Docetaksel	0(,0)	0(,0)	4(3,7)	0(,0)	0(,0)	0(,0)	0(,0)			
	ADT+ARPİ	20(80,0)	0(,0)	64(58,7)	9(15,8)	0(,0)	0(,0)	0(,0)			
	ADT+ARPİ+Docetaksel	3(12,0)	0(,0)	40(36,7)	48(84,2)	0(,0)	0(,0)	0(,0)			
	ADT+Lokal Tedavi	0(,0)	0(,0)	0(,0)	0(,0)	0(,0)	0(,0)	0(,0)			

	İzlem	0(,0)	0(,0)	0(,0)	0(,0)	0(,0)	0(,0)	0(,0)			
	Diğer	0(,0)	0(,0)	0(,0)	0(,0)	0(,0)	0(,0)	0(,0)			
Claude	ADT	8(32,0)	0(,0)	0(,0)	0(,0)	0(,0)	0(,0)	0(,0)	0,532	<0,001	73,6
	ADT+Docetaksel	0(,0)	0(,0)	0(,0)	0(,0)	0(,0)	0(,0)	0(,0)			
	ADT+ARPİ	16(64,0)	0(,0)	88(80,7)	11(19,3)	0(,0)	0(,0)	0(,0)			
	ADT+ARPİ+Docetaksel	0(,0)	0(,0)	20(18,3)	46(80,7)	0(,0)	0(,0)	0(,0)			
	ADT+Lokal Tedavi	1(4,0)	0(,0)	1(,9)	0(,0)	0(,0)	0(,0)	0(,0)			
	İzlem	0(,0)	0(,0)	0(,0)	0(,0)	0(,0)	0(,0)	0(,0)			
	Diğer	0(,0)	0(,0)	0(,0)	0(,0)	0(,0)	0(,0)	0(,0)			
Gemini	ADT	6(24,0)	0(,0)	1(,9)	0(,0)	0(,0)	0(,0)	0(,0)	0,412	<0,001	65,8
	ADT+Docetaksel	0(,0)	0(,0)	0(,0)	0(,0)	0(,0)	0(,0)	0(,0)			
	ADT+ARPİ	16(64,0)	0(,0)	73(67,0)	9(15,8)	0(,0)	0(,0)	0(,0)			
	ADT+ARPİ+Docetaksel	3(12,0)	0(,0)	34(31,2)	48(84,2)	0(,0)	0(,0)	0(,0)			
	ADT+Lokal Tedavi	0(,0)	0(,0)	1(,9)	0(,0)	0(,0)	0(,0)	0(,0)			
	İzlem	0(,0)	0(,0)	0(,0)	0(,0)	0(,0)	0(,0)	0(,0)			
	Diğer	0(,0)	0(,0)	0(,0)	0(,0)	0(,0)	0(,0)	0(,0)			
Deep Seek	ADT	2(7,7)	0(,0)	0(,0)	0(,0)	0(,0)	0(,0)	0(,0)	0,150	0,003	58,5
	ADT+Docetaksel	0(,0)	0(,0)	0(,0)	2(3,5)	0(,0)	0(,0)	0(,0)			
	ADT+ARPİ	24(92,3)	0(,0)	94(85,5)	38(66,7)	0(,0)	0(,0)	0(,0)			
	ADT+ARPİ+Docetaksel	0(,0)	0(,0)	16(14,5)	17(29,8)	0(,0)	0(,0)	0(,0)			
	ADT+Lokal Tedavi	0(,0)	0(,0)	0(,0)	0(,0)	0(,0)	0(,0)	0(,0)			

	İzlem	0(,0)	0(,0)	0(,0)	0(,0)	0(,0)	0(,0)	0(,0)			
	Diğer	0(,0)	0(,0)	0(,0)	0(,0)	0(,0)	0(,0)	0(,0)			