Big Data Meets Public Health: A Logistic Regression Analysis of Vitamin D Deficiency in the U.S. using the National Institutes of Health's All of Us Database

Mayra S. Haedo Cruz
Master in Computer Science
Advisor: Dr. Luis M. Vicente
Polytechnic University of Puerto Rico
Graduate Project EXPO, May 2025

Abstract *— Vitamin D deficiency is a long-standing public health issue associated with various chronic conditions. This study explores how genetic ancestry, specific single nucleotide polymorphisms (SNPs), and solar radiation exposure influence the risk of vitamin D deficiency in the diverse U.S. population using data from the large-scale All of Us project. With a matched case-control study of 16,145 vitamin D-deficient and 16,145 vitamin D-sufficient participants, logistic regression was used to assess these associations. Findings reveal that African ancestry, particular SNP variants, and lower solar radiation exposure are significant predictors of deficiency. The results also show that risk factors differ among racial groups, emphasizing the complexity of gene-environment interactions. This research contributes to a deeper understanding of the biological and environmental drivers of vitamin D deficiency and may support the development of personalized and population-specific public health strategies to address disparities in vitamin D-related health outcomes.*

Keywords *— Genetic Ancestry, Logistic Regression, SNPs, Vitamin D Deficiency.*

INTRODUCTION

Vitamin D deficiency remains a critical public health concern in the United States, with far-reaching implications for population health and health equity [1] [2] [3]. Traditionally known for its role in calcium homeostasis and skeletal integrity, vitamin D is now recognized as a pleiotropic hormone involved in a wide range of physiological processes, including immune regulation, inflammation control, and cellular proliferation. Deficiency in vitamin D has been associated with an increased risk for numerous chronic conditions, including osteoporosis, autoimmune diseases, cardiovascular disease, certain cancers, and neurodegenerative disorders such as Alzheimer's disease [4]. Despite these known associations, the prevalence of vitamin D deficiency remains high, particularly among racially and ethnically minoritized populations [3] [5] [6].

While previous studies have explored genetic and environmental contributors to vitamin D deficiency, these factors are often examined in isolation. Such approaches fail to capture the complex interplay between an individual's genetic background and their lived environment. This study addresses that critical gap by integrating three key domains: genetic ancestry, specific vitamin D-related single nucleotide polymorphisms (SNPs), and high-resolution environmental solar radiation data. Using data from the All of Us research program, a large, diverse, and nationally representative cohort, we employ logistic regression modeling to identify predictors of deficiency and assess how these risk factors vary across genetic ancestries and geographic contexts.

Paradoxically, high rates of vitamin D deficiency are frequently observed in regions with abundant sunshine [7]. This apparent contradiction highlights the complexity of vitamin D synthesis, which depends not only on ultraviolet (UV) radiation exposure but also on behavioral, environmental, and physiological factors. Individuals may live in sunny environments but occupy them for short periods, use UV-blocking sunscreen, or live in areas with high air pollution levels that reduce effective UVB penetration [1] [7]. Furthermore, skin pigmentation plays a significant role in cutaneous vitamin D production. Melanin protects the skin against UV

radiation and thus reduces the skin's capacity to synthesize vitamin D. Consequently, individuals with darker skin tones, predominantly those of African, Caribbean, or Latin American descent, require more sun exposure to achieve sufficient vitamin D levels [6] [8]. This physiological difference, coupled with inequities in healthcare access, education, and nutrition, places these populations at greater risk for vitamin D deficiency and its associated health consequences.

The concept of race further complicates the study of vitamin D deficiency. Although race is a social construct with no biological basis, it remains a powerful determinant of health due to its influence on lived experience, access to care, exposure to discrimination, and socioeconomic status [6] [8]. Incorporating genetic ancestry into our analysis allows for a more nuanced understanding of biological predisposition while acknowledging that socially defined racial categories still reflect significant disparities in environmental exposures and health outcomes [8]. This dual perspective is essential in addressing disparities rooted in both structural and biological domains. Considering that racial disparities impact health, we are stratifying our analyses based on self-reported race to understand how these subpopulations are impacted differently by these factors.

Beyond genetic ancestry, several single nucleotide polymorphisms (SNPs) have been consistently associated with circulating vitamin D levels across diverse populations. Among these, rs2282679 located in the GC gene (encoding the vitamin D-binding protein), rs12785878 near the DHCR7 gene (involved in cholesterol synthesis and precursor availability for vitamin D production), and rs10741657 in the CYP2R1 gene (which encodes a key enzyme for 25-hydroxylation of vitamin D) are the most extensively studied. These SNPs influence vitamin D metabolism, transport, and activation, and have been linked to both baseline serum 25(OH)D concentrations and response to supplementation. Understanding the distribution and impact of these variants across different ancestral backgrounds is crucial for elucidating genetic contributions to vitamin D deficiency and informing precision nutrition and public health strategies.

This is one of the largest and most comprehensive analyses to date that examines vitamin D deficiency through the combined lens of genetic ancestry, SNP variation, and environmental UV exposure. By leveraging the scale and diversity of the All of Us database, this study provides novel insights into the biological and social mechanisms underlying vitamin D deficiency. Ultimately, our findings aim to inform targeted, equity-oriented public health interventions that account for the complex, intersectional nature of health disparities in the United States.

METHODS

This section will observe the cohort design and analysis approach for the logistic regression modeling.

Cohort

- Be 18 years old or older at the age of consent.
- Have zip code data.
- Have Infinium Global Diversity Array data.
- Have sex at birth data.
- Have laboratory data for blood serum levels for 25-hydroxyvitamin D_3 in nanomoles/milliliters.
- Have blood serum levels of 25-hydroxyvitamin D_3 between 0 nm/mL and 50nm/mL.
- Have data for the three SNPs that have been identified to be related to vitamin D deficiency: rs2282679, rs12785878, and rs10741657.
- Have data for the date of blood serum collection.

Total cohort: 16,145 cases, 16,145 controls

Analysis Approach

For individuals with multiple records of serum 25-hydroxyvitamin D_3 (25(OH)D), only the earliest measurement will be retained for analysis, while any subsequent measurements will be excluded to maintain consistency. Participants with serum 25(OH)D levels ranging from 0 to 20.4 ng/mL will be categorized as "deficient," while those with levels between 20.5 and 50 ng/mL will be considered

"normal." Measurements exceeding 50 ng/mL will be excluded from analysis due to the potential for vitamin D toxicity and the lack of standardized clinical interpretation beyond this range. Participants missing race or ethnicity data will be categorized as "Not Specified" to preserve sample size and minimize bias introduced by listwise deletion.

Each deficient participant will be matched to a non-deficient control based on age, sex assigned at birth, self-reported race and ethnicity, and detailed genetic ancestry components, including African, European, American Indigenous, East Asian, South Asian, and Middle Eastern ancestry proportions. This matching approach is designed to reduce confounding and enhance the precision of comparative analyses across diverse backgrounds.

Geographic location will be inferred using each participant's residential zip code, which will be mapped to the corresponding state using United States Postal Service (USPS) zip-code directories. Annual average solar radiation data for each U.S. state, including Alaska and Puerto Rico, will be incorporated as a continuous environmental variable from the Centers for Disease Control's National Environmental Public Health Tracking Network, which offers publicly available geospatial solar radiation datasets.

Genotypic data for key vitamin D-related single nucleotide polymorphisms (SNPs) will be extracted from genome-wide microarray data using PLINK, a standard tool in genetic epidemiology. SNPs will be selected based on prior evidence of association with vitamin D metabolism, transport, and receptor activity.

To assess the relative contribution of genetic, environmental, and demographic factors to vitamin D deficiency, a series of logistic regression models will be constructed using a generalized linear model (GLM) framework [9]. An initial model will be fitted for the general U.S. population to estimate overall predictors of deficiency risk. Stratified models will then be constructed separately for three major racial and ethnic groups, Black or African American, White or European American, and Latino or Hispanic participants, to assess whether risk factors vary across ancestries and racial identities. These subgroup-specific GLMs will help uncover unique genetic or environmental vulnerabilities that may be masked in aggregated population-level analyses, offering a more granular understanding of disparities in vitamin D deficiency risk. The models will include fixed effects for SNP genotypes, ancestry percentages, solar radiation, and sociodemographic variables.

RESULTS

The following section will demonstrate the logistic regression results for the general cohorts, and then stratified by subpopulation.

General U.S. Population

In the following results, we observe the odds ratio, which represent the impact that different predictor variables have regarding vitamin D deficiency risk in the general U.S. population:

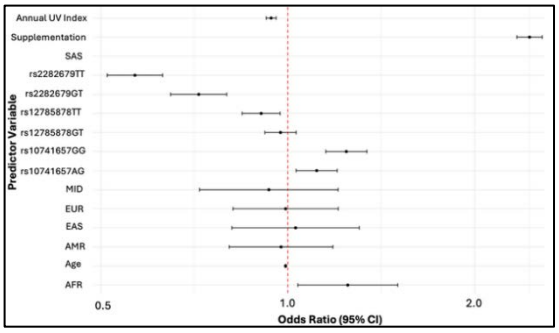

Figure 1
Forest Plot of Logistic Regression Coefficients for the General U.S. Population

Table 1
Odds Ratio and Confidence Intervals for Vitamin D Risk in the General US Population

Variable	OR*	p-value	CI**
UV Index	0.9402	6.56×10^{-11}	0.92-0.96
Supplement	2.451	1.19×10^{-305}	2.3-2.6
Age	0.9919	9.58×10^{-25}	0.99-0.99
SAS (South Asian)	-	-	-
MID (Middle Eastern)	0.9323	5.92×10^{-1}	0.72-1.2
EUR (European)	0.9920	9.36×10^{-1}	0.81-1.2
EAS (East Asian)	1.029	8.10×10^{-1}	0.81-1.3
AMR (American Indigenous and Native)	0.9755	8.00×10^{-1}	0.80-1.2

AFR (African and Black)	1.249	$1.85x10^{-2}$	1.1-1.5
rs2282679TT	0.5674	$3.21x10^{-27}$	0.51-0.63
rs2282679GT	0.7189	$5.28x10^{-10}$	0.65-0.80
rs12785878TT	0.9061	$6.01x10^{-3}$	0.84-0.97
rs12785878GT	0.9734	$3.65x10^{-1}$	0.92-1.0
rs10741657GG	1.243	$5.77x10^{-3}$	1.2-1.3
rs10741657AG	1.113	$2.32x10^{-8}$	1.0-1.2

*OR=Odds Ratio, **CI=Confidence Interval

Table 1 represents the numeric values used to construct Figure 1. In the general U.S. population, logistic regression analysis identified several significant predictors of vitamin D deficiency (Figure 1, Table 1). Vitamin D supplementation emerged as the strongest protective factor, which is consistent with previous studies considering that this is the primary method for treating vitamin D deficiency (OR = 2.451, 95% CI: 2.3–2.6, p = 1.19×10^{-305}) We will see this same result in the following stratified analyses. Similarly, higher annual average UV Index was associated with lower odds of deficiency by ~6% per UV level increase (OR = 0.9402, 95% CI: 0.92–0.96, p = 6.56×10^{-11}), emphasizing the critical role of environmental sunlight exposure in maintaining sufficient vitamin D levels. Age was also inversely associated with deficiency, although the effect size was small (OR = 0.9919, 95% CI: 0.99–0.99, p = 9.58×10^{-25}).

Among the ancestry variables, African ancestry showed a significant trend toward increased odds of vitamin D deficiency, where one percent of increased African ancestry augmented the chances for vitamin D deficiency by ~25% (OR = 1.249, 95% CI: 1.1-1.5, p = 0.0185). Other ancestry groups, including European, East Asian, American Indigenous, and Middle Eastern, did not exhibit significant associations in the general model. There was not enough data of participants with South Asian ancestry to obtain results for this ancestry component.

Genetic analysis of vitamin D-related SNPs revealed several significant associations. The rs2282679TT and rs2282679GT genotypes were strongly associated with reduced odds of deficiency (OR = 0.5674, 95% CI: 0.51–0.63, p = 3.21×10^{-27}; OR = 0.7189, 95% CI: 0.65–0.80, p = 5.28×10^{-10}, respectively). Similarly, carriers of the rs12785878TT genotype had reduced odds of deficiency by ~9% (OR = 0.9061, 95% CI: 0.84–0.99, p = 6.01×10^{-3}). Conversely, the rs10741657GG and rs10741657AG genotypes were associated with increased risk (OR = 1.243, 95% CI: 1.2–1.3, p = 5.77×10^{-3}; OR = 1.113, 95% CI: 1.0–1.2, p = 2.32×10^{-8}, respectively). These findings highlight the significant contribution of specific genetic variants to vitamin D deficiency risk, independent of environmental factors.

Hispanic or Latino Population

In the stratified analysis restricted to Latino participants (n = 7,980), several demographic, environmental, and genetic predictors were significantly associated with vitamin D deficiency:

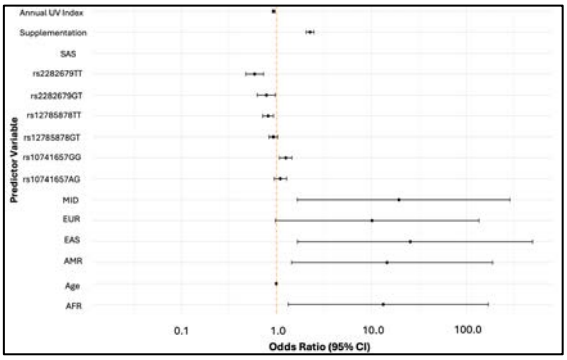

Figure 2
Forest Plot of Logistic Regression Coefficients for the Hispanic or Latino Population

Annual average UV Index remained a protective factor, with higher environmental UV exposure associated with lower odds of deficiency (~7% per unit increase, OR = 0.93, 95% CI: 0.90–0.96, p < 0.0001). Age also showed a small but statistically significant inverse association with deficiency risk (OR = 0.99, 95% CI: 0.99–1.00, p < 0.0001), though the effect size was minimal.

Within this subgroup, higher proportions of East Asian, American Indigenous, Middle Eastern, and African genetic ancestry were each independently associated with increased odds of vitamin D deficiency. The most pronounced association was observed for East Asian ancestry (OR = 25.64, 95% CI: 1.65–497.97, p = 0.024), though this estimate was accompanied by wide confidence intervals,

suggesting variability or limited representation of this ancestry component in the cohort. Similarly, American Indigenous (OR = 14.52, 95% CI: 1.43–189.15, p = 0.031), Middle Eastern (OR = 19.44, 95% CI: 1.64–288.73, p = 0.023), and African (OR = 13.24, 95% CI: 1.31–170.04, p = 0.032) ancestries were associated with higher deficiency risk, highlighting potential ancestral disparities within the Latino population, specifically considering they are a very admixed population. There was not enough data of participants with South Asian ancestry to obtain results for this ancestry component.

Genotypic analyses also revealed significant associations. The rs2282679TT genotype exhibited a strong protective effect with carries having reduced odds of deficiency by ~42% (OR = 0.58, 95% CI: 0.47–0.73, p < 0.0001), while rs2282679GT also conferred lower odds of deficiency with carries having reduced odds of deficiency by ~22% (OR = 0.78, 95% CI: 0.62–0.97, p = 0.028). Conversely, carriers of the rs10741657GG genotype had higher odds of deficiency (OR = 1.24, 95% CI: 1.07–1.44, p = 0.004), and the rs12785878TT genotype was associated with decreased odds of deficiency (OR = 0.81, 95% CI: 0.71–0.93, p = 0.0018).

African American or Black Population

In the stratified analysis focusing exclusively on Black participants (n = 8,947), logistic regression analysis revealed several significant predictors of vitamin D deficiency:

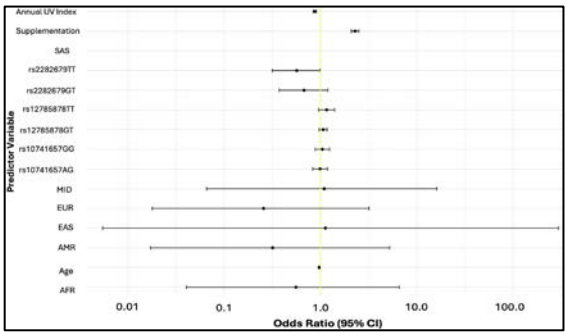

Figure 3
Forest Plot of Logistic Regression Coefficients for the African American or Black Population

Again, higher annual UV Index was associated with significantly lower odds of deficiency (OR = 0.87, 95% CI: 0.84–0.91, p < 0.0001), reinforcing the protective role of sunlight exposure within this subgroup. Age also demonstrated a small but significant inverse association with deficiency (OR = 0.97, 95% CI: 0.96–0.97, p < 0.0001), consistent with trends observed in other population strata.

Ancestry components did not display significant associations in this model, including African ancestry proportion itself (OR = 0.56, 95% CI: 0.04–6.60, p = 0.64). The wide confidence intervals observed for European, American Indigenous, and Middle Eastern ancestry estimates reflect limited variability and lower representation of these ancestries in the Black participant cohort. There was not enough data of participants with South Asian ancestry to obtain results for this ancestry component.

Regarding genetic variants, rs2282679TT and rs2282679GT genotypes were both associated with lower odds of deficiency (OR = 0.57, 95% CI: 0.31–0.99, p = 0.052; OR = 0.67, 95% CI: 0.37–1.19, p = 0.18, respectively), although statistical significance was borderline or not reached.

White Population

In the stratified analysis of White participants (n = 12,791), several significant predictors of vitamin D deficiency were identified:

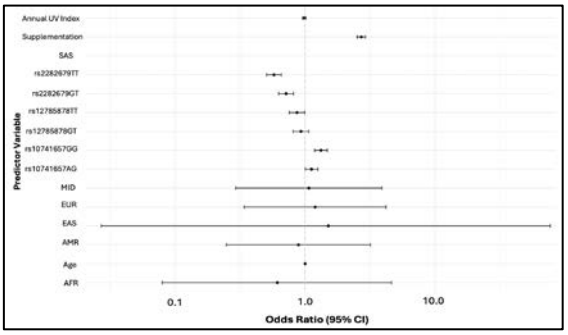

Figure 4
Forest Plot of Logistic Regression Coefficients for the White Population

Interestingly, in contrast to other groups, age exhibited a small yet statistically significant positive association with deficiency risk (OR = 1.00, 95% CI: 1.001–1.006, p = 0.0029), although the effect size was minimal. Annual UV Index was not

significantly associated with deficiency in this subgroup (OR = 0.99, 95% CI: 0.96–1.02, p = 0.44), suggesting that environmental UV exposure might play a less pronounced role in vitamin D status among White participants compared to other groups.

Among genetic variables, the rs2282679TT genotype exhibited a strong protective effect (OR = 0.58, 95% CI: 0.51–0.66, p < 0.0001), consistent with findings in other ancestry groups. Carriers of the rs2282679GT genotype also had reduced odds of deficiency (OR = 0.72, 95% CI: 0.63–0.82, p < 0.0001). Similarly, the rs10741657GG genotype was significantly associated with increased odds of deficiency (OR = 1.33, 95% CI: 1.18–1.49, p < 0.0001), while rs10741657AG also demonstrated a smaller but statistically significant association (OR = 1.12, 95% CI: 1.01–1.26, p = 0.04). The rs12785878TT genotype was associated with a modest protective effect (OR = 0.87, 95% CI: 0.76–0.99, p = 0.045), whereas the rs12785878GT genotype was not significantly associated. Ancestry components, including proportions of African, East Asian, American Indigenous, and Middle Eastern ancestries, were not significant predictors within the White participant model. There was not enough data of participants with South Asian ancestry to obtain results for this ancestry component.

DISCUSSION

Consistently across all models, vitamin D supplementation emerged as the most robust protective factor, with the strongest effect observed among White participants. This finding underscores the effectiveness of supplementation as a direct intervention to address deficiency regardless of genetic or environmental background. However, our study also highlights that reliance on supplementation alone may not address the underlying disparities rooted in ancestry, genetics, and environment, particularly in historically underserved populations.

Environmental UV exposure demonstrated varying degrees of influence depending on the population. In the general U.S., Latino, and Black populations, UV exposure was significantly associated with reduced odds of deficiency, reaffirming the role of sunlight in endogenous vitamin D synthesis. Interestingly, this association was not significant among White participants, suggesting that behavioral or cultural factors, rather than environmental UV availability alone, might play a larger role in determining vitamin D status in this group.

The Latino population, characterized by high genetic admixture, showed strong and significant associations between ancestry proportions, particularly East Asian, African, American Indigenous, and Middle Eastern ancestries, and increased risk of deficiency. These findings underscore the importance of considering the unique genetic and social contexts of admixed populations when addressing vitamin D disparities. This is especially important when we consider that different ethnic populations within Latinos (Puerto Ricans, Mexicans, Pervuians, Cubans, Venezuelans, etc.) have, on average, different distributions of genetic ancestry percentages. In further studies, we could stratify these results based on specific ethnic background within Latino populations in the All of Us database.

Genetic variants demonstrated consistent associations across all groups. The rs2282679TT and rs2282679GT genotypes consistently showed protective associations, while the rs10741657GG and rs10741657AG genotypes were associated with increased odds of deficiency, particularly in the general and White populations. The strength and consistency of these SNP associations across racial and ethnic groups support their role as important genetic markers for vitamin D deficiency risk, independent of environmental exposures or demographic factors. However, among Black participants, some associations did not reach statistical significance, possibly reflecting the lower variability of these SNPs within this population or the need for larger sample sizes to detect effects.

CONCLUSION

Overall, our findings demonstrate that while supplementation remains a universal and potent intervention for vitamin D deficiency, the pathways leading to deficiency are multifaceted and vary across racial and ethnic groups. Ancestry, genetic predisposition, and environmental exposures all contribute to individual risk profiles, and their effects differ in magnitude and direction depending on the population context. This suggests that public health interventions aiming to reduce vitamin D deficiency disparities must be tailored, considering both genetic and sociocultural factors that uniquely impact each group.

Our study highlights the need for precision public health strategies that integrate genomics, environmental data, and social determinants of health to more effectively address vitamin D deficiency. By doing so, we can move beyond one-size-fits-all recommendations and design interventions that are equitable, culturally appropriate, and biologically informed. Future studies include the validation of these findings in longitudinal cohorts, random forest modelling of this disease, and exploring gene-environment interactions more deeply, particularly within admixed populations such as Latinos and individuals of African descent, where current gaps in representation and data persist.

REFERENCES

[1] K. Y. Forrest, et al., "Prevalence and correlates of vitamin D deficiency in US adults," in *Nutrition Research*, vol. 1, pp. 48-54, Jan. 31, 2011. DOI: 10.1016/j.nutres.2010.12.001.

[2] A. C. Looker, et al., "Vitamin D status: United States, 2001–2006," in *NCHS Data Brief*, no. 59, March 2011. Available: https://www.cdc.gov/nchs/data/databriefs/db59.pdf.

[3] F. Bandeira, et al., "Vitamin D deficiency: a global perspective," in *Arquivos Brasileiros de Endocrinologia & Metabologia*, vol. 50, no. 4, Aug. 2006. DOI: https://doi.org/10.1590/S0004-27302006000400009.

[4] M. S. LeBoff, et al., "Effects of Supplemental Vitamin D on Bone Health Outcomes in Women and Men in the VITamin D and OmegA-3 TriaL (VITAL)," in *Journal of Bone and Mineral Research*, vol. 35, no. 5, pp. 883-893, Jan. 30, 2020. DOI: 10.1002/jbmr.3958.

[5] A. A. Ginde, et al., "Demographic differences and trends of vitamin D insufficiency in the US population, 1988–2004," in *Archives of Internal Medicine*, vol. 169, no. 6, pp. 626-632, Mar. 23, 2009. DOI: 10.1001/archinternmed.2008.604.

[6] S. Yao, et al., "Demographic, lifestyle, and genetic determinants of circulating concentrations of 25-hydroxyvitamin D and vitamin D-binding protein in African American and European American women," in *The American Journal of Clinical Nutrition*, vol. 105, no. 6, pp. 1362-1371, Jun. 2017. DOI: 10.3945/ajcn.116.143248.

[7] N. Binkley, et al., "Low vitamin D status despite abundant sun exposure," in *The Journal of Clinical Endocrinology & Metabolism*, vol. 92, no. 6, Jun. 1, 2007. DOI: 10.1210/jc.2006-2250.

[8] L. B. Signorello, et al., "Race, genetic ancestry, and vitamin D deficiency," in *Cancer Epidemiology, Biomarkers & Prevention*, vol. 105, no. 12, pp. e4337-e4350, Dec. 1, 2020. DOI: 10.1210/clinem/dgaa612.

[9] M. Meysami, et al., "Utilizing logistic regression to compare risk factors in disease modeling with imbalanced data: a case study in vitamin D and cancer incidence," in *Frontiers in Oncology*, vol. 13, Sept. 28, 2023. DOI: 10.3389/fonc.2023.1227842.