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[1,2], Julie Dutil[1], and Josué Pérez-Santiago[1,3]

Advisor: Dr. Luis Vicente[2]

[1]Division of Clinical and Translational Cancer Research, University of Puerto Rico Comprehensive Cancer Center, San Juan, Puerto Rico;
[2]Polytechnic University of Puerto Rico, San Juan, Puerto Rico;
[3]University of Puerto Rico, Medical Sciences Campus, San Juan, Puerto Rico

Mayra S. Haedo-Cruz, B.Sc.
UPR Comprehensive Cancer Center
E-mail: mayra.haedo@upr.edu

Josué Pérez Santiago, Ph.D.
UPR Comprehensive Cancer Center
E-mail: josue.perez22@upr.edu
Phone: (787) 772-8300 x1219

JPS Lab Linktree: M. Haedo Linkedin:

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 in the general U.S. population. The results also show that risk factors differ among subgroups, 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.

Introduction

- Vitamin D deficiency is a known ailment to cause **increased risk of excess mortality, infections, and other diseases**.
- Despite widespread availability of sunlight and supplementation, **deficiency remains prevalent across various populations in the US**, particularly in individuals with higher melanin levels, limited sun exposure, or genetic predispositions.
- **Understanding the interaction** between genetic variants, ancestry, and environmental factors like location-based UV-index is essential for addressing disparities in vitamin D status.

Background

- **Key SNPs**: Single nucleotide polymorphisms (SNPs) like rs2282679 (in **GC**), rs12785878 (near **DHCR7**), and rs10741657 (in **CYP2R1**) are common genetic variations that influence how vitamin D is transported, synthesized, and activated in the body.
- **Race vs. Ancestry**: While race is a social construct, it impacts health through structural inequalities. Including genetic ancestry helps separate social influences from biological factors in vitamin D deficiency risk.
- **Sunshine Paradox**: Deficiency can occur even in sunny areas due to low sun exposure, sunscreen use, pollution, and melanin-rich skin, which reduces vitamin D production.

Problem

- Determine the magnitude of the drivers of Vitamin D deficiency in the US population: **Genetic ancestry, single nucleotide polymorphisms, or sun exposure.**
- **No study to date** has integrated detailed genetic profiles, including multiple ancestry components and key SNPs, with precise, high-resolution UV exposure data derived from zip-code level solar radiation measures for US participants and its various subpopulations at this scale.

Methodology

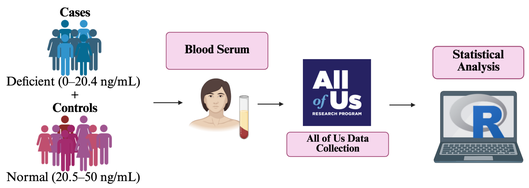


Figure 1. Methods. Individuals within the All of Us (AoU) database were matched on age, sex, race, ethnicity, and ancestry proportions. A logistic regression model assessed the relationship between vitamin D status and age, sex, UV radiation, ancestry, and genotype for SNPs rs2282679, rs12785878, and rs10741657. Analysis was repeated by subpopulations. RStudio and Python was used.

Results

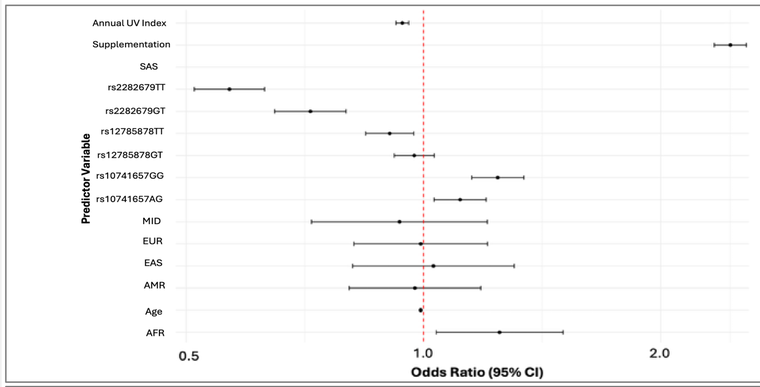
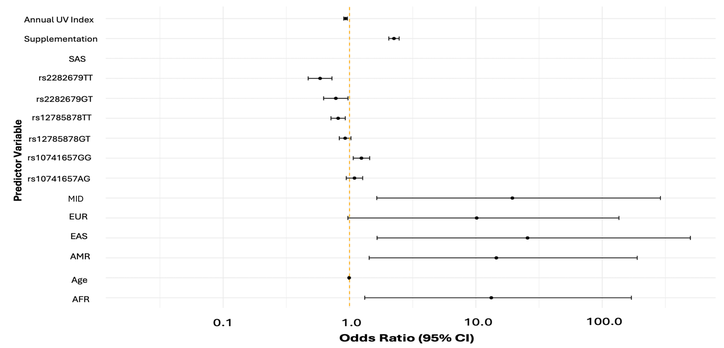


Figure 1. Forest Plot of Logistic Regression Odds Ratios in the General U.S. Population. Higher annual average UV Index and older age were associated with lower odds of deficiency, highlighting the importance of sun exposure and age-related changes in vitamin D metabolism. African ancestry was linked to a significantly higher risk of deficiency (~25% increase per unit, OR = 1.25, 95% CI: 1.1–1.5), while other ancestry groups showed no significant associations. Several genetic variants demonstrated strong effects: rs2282679TT and rs12785878TT genotypes were protective, while rs10741657GG and AG were associated with increased risk.

Figure 2. Forest Plot of Logistic Regression Odds Ratios in the Hispanic or Latino Population. Higher UV Index and older age were both significantly associated with reduced odds of deficiency. Several ancestry components including East Asian, American Indigenous, Middle Eastern, and African were significantly associated with elevated risk, with East Asian ancestry showing the largest estimated effect (OR = 25.64). Genotypic analysis showed that rs2282679TT and GT were protective, while rs10741657GG increased risk, and rs12785878TT was also associated with lower odds of deficiency.



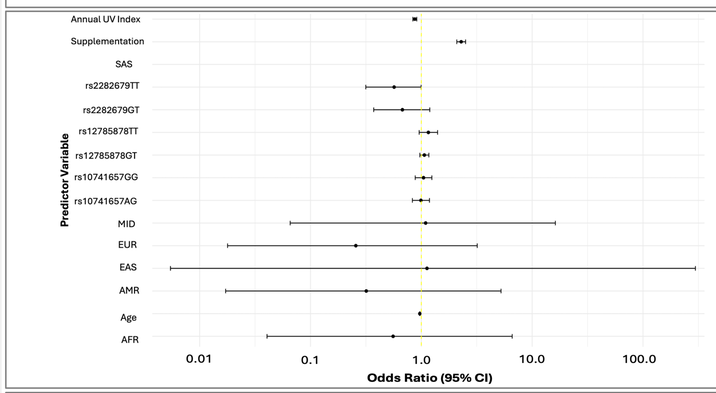
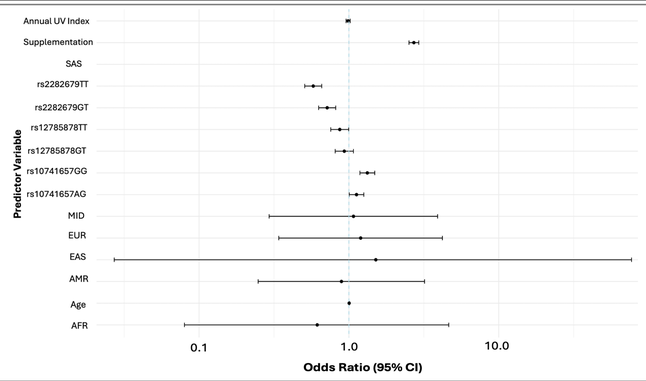


Figure 3. Forest Plot of Logistic Regression Odds Ratios in the Black or African American Population. A higher annual UV Index remained a significant protective factor, with each unit increase associated with ~13% lower odds of deficiency (OR = 0.87, 95% CI: 0.84–0.91). Age was also inversely associated with deficiency, although the effect was modest. Ancestry components did not reach statistical significance. Genetic analysis revealed that rs2282679TT showed a potential protective trend (OR = 0.57), though statistical significance was borderline (p = 0.052).

Figure 4. Forest Plot of Logistic Regression Odds Ratios in the White Population. Unlike other subgroups, age showed a small but statistically significant positive association with deficiency risk, while UV Index was not a significant predictor, suggesting a reduced role of environmental sunlight exposure in this population. Several genetic variants emerged as key predictors: rs2282679TT and GT genotypes were strongly protective (OR = 0.58 and 0.72, respectively), whereas rs10741657GG and AG were associated with increased risk. The rs12785878TT genotype also conferred modest protection (OR = 0.87), though rs12785878GT was not significant. Ancestry proportions did not significantly influence deficiency risk in this model.



Discussion and Conclusions

- **Supplementation as a universal benefit:** Across all groups, vitamin D supplementation was the strongest protective factor, especially among White participants, but it may not fully address disparities rooted in ancestry, environment, or systemic inequities.
- **UV exposure varies by group:** Environmental UV exposure significantly reduced deficiency risk in Latino, Black, and general populations, but not in White participants, pointing to potential behavioral or cultural influences.
- **Ancestry matters in admixed groups:** In Latinos, higher proportions of East Asian, African, American Indigenous, and Middle Eastern ancestry were linked to greater deficiency risk, emphasizing the importance of accounting for genetic admixture in public health strategies.
- **Consistent genetic effects:** Genotypes such as rs2282679TT (protective) and rs10741657GG (risk-increasing) were associated with vitamin D status across most groups, though associations were weaker in Black participants.
- **Addressing vitamin D deficiency effectively requires integrating genomics, environmental exposure, and social determinants to develop targeted, culturally informed interventions, especially for underserved and admixed populations.**

Future Directions

- Perform **ANOVA tests** to assess distribution between SNP genotypes and ancestries.
- Replicate this model solely on individuals who **do not supplement** for vitamin D deficiency.
- Perform a **Random Forest** analysis to predict likelihood of vitamin D deficiency.
- Incorporate **more factors** such as BMI, education level, income level, and family relationships for the models.
- **Compare the outcome** of the logistic regression model in the general U.S. population vs. the factor importance of the random forest on the general U.S. population.

References

Acknowledgements

This work was in part funded by the AIM-AHEAD All of Us Training Program. We gratefully acknowledge All of Us participants for their contributions, without whom this research would not have been possible. We also thank the National Institutes of Health's All of Us Research Program for making available the participant data examined in this study.