AARHUS UNIVERSITY

regionmidtjylland

Master's thesis in Bioinformatics

Sex Chromosome Trisomies and Cardiovascular Disease in a Genetically Diverse Population

A Genotype-first screening of the *All of Us Program* Cohort

Author:

Unai López-Errasti

Student Number: 202402138

Supervisors:

Jonas Ghouse & Juraj Bergman

Scm

MSc in Bioinformatics

Aarhus University

June 2026

Acknowledgements

The way I see it, the Master's thesis is something to celebrate, a celebration of the transition from the controlled world to the wild one. An important milestone that, beyond representing all the hard work carried out during these two years, acts as a departure point for even more ambitious goals.

This project was an incredible opportunity to learn how science is done, the questions one might ask oneself, and to formulate questions that truly make the difference. This journey also taught me that in science and research, things go unexpectedly more often than one might think, with downs that might take you down. However, knowing how to deal with these situations and trying to see the positive side so you can keep pushing, even if that means starting again, is something I will never forget.

First and foremost, I want to thank my family, my girlfriend, and my friends for being the place to disconnect when life makes you feel overwhelmed.

A huge thank you to Juraj, for his continued guidance, help, and involvement throughout the process.

Another huge thank you to Jonas, who gave me the incredible opportunity to put into practice what I have been learning during these years, who shared his knowledge with me, and who encouraged me to think rather than follow instructions like an AI model. I am looking forward to continuing to work with him in the years to come.

Finally, I also want to thank the National Institutes of Health's All of Us Research Program and the All of Us participants for making available the data examined in this study.

Generative AI usage declaration

I have used generative artificial intelligence (GAI) tools in the completion of this project.

Tools Used:

- Claude (Version Sonnet 4.6, 5 and Opus 4.6, 4.7, 4.8) by Anthropic
- Gemini (Version 3 and 3.1) by Alphabet
- LFM2 (Version 24B A2B, Local) by Liquid AI

Use Cases:

- **For programming tasks:** I used Claude as a support tool for Python code debugging. I used Claude to generate and refine SQL queries to retrieve data from the All of Us Curated Data Repository via gstore.
- **For feedback on own text and contrast of ideas:** I used Gemini and Claude to refine both the writing style and the organization of the report and, to correct typos. I also used it as a tool to compare different technical possibilities and "decision-making" paths throughout the research process.
- **To aid in setting up the Latex document:** I used Gemini to troubleshoot formatting issues, improve document structure, and implement specific packages for citation styles, table layout, and figure placement.

In all cases, I have critically and carefully reviewed all GAI-generated outputs included in this thesis. The resulting analysis, interpretations, and conclusions are my own, and I remain solely responsible for the entire content of this document.

Abstract

Importance

Sex chromosome aneuploidies (SCAs) are often undiagnosed. Their cardiovascular and cardiometabolic burden across genetically diverse populations is poorly characterized.

Objective

To determine whether genomically identified SCAs are associated with cardiovascular and cardiometabolic disease using survival analysis and whether clinical recognition differs by genetic ancestry.

Design, Setting, and Participants

Retrospective cohort study of SCAs and cardiometabolic disease, performed by analyzing X- and Y-chromosome dosage from Global Diversity Array intensity data in 422,818 participants of the All of Us Research Program (CDR v9). Ancestry was inferred genetically, with 41.6% of participants of non-European ancestry, permitting ancestry-stratified analysis not possible in previous cohorts. A total of 570 carriers of 47,XXY, 47,XYY or 47,XXX were identified and linked to electronic health records, with a median follow-up of 9.7 years (IQR 4.9–16.0).

Exposure

Genetically identified sex chromosome trisomies: 47,XXY, 47,XYY, or 47,XXX.

Main Outcomes and Measures

Composite cardiovascular disease (merged ICD diagnosis of 8 different cardiac diseases) and secondary cardiometabolic outcomes (ICD diagnosis of 11 distinct cardiometabolic diseases). Documented ICD diagnosis of sex SCA among genetically identified carriers.

Results

Sex chromosome aneuploidy was identified in 570 of 422,818 participants (134.81 per 100,000, 95% CI 123.97–146.34), comprising 253 47,XXY, 142 47,XYY and 175 47,XXX carriers. Only 77 carriers (13.5%) had a corresponding clinical diagnosis recorded, ranging from 22.52% for

47,XXY to below 11.42% for 47,XXX. A total of 555 carriers without prevalent disease were included in time-to-event analyses. Carriers had almost twice the hazard of the composite cardiovascular outcome (hazard ratio, 1.92 [95% CI, 1.60–2.31]; P = 3.23E-12; 114 events), with 10-year cumulative incidence of 16% compared with 11% in non-carriers. Eight of eleven individual cardiometabolic outcomes were significantly associated after false discovery rate correction, to highlight aortic stenosis (hazard ratio, 3.52 [95% CI, 2.04–6.07]) and venous thromboembolism (hazard ratio, 3.11 [95% CI, 2.22-4.36]). Aneuploidy prevalence was higher in African (164.53 per 100,000 [95% CI, 137.28–195.60]) and American (162.69 per 100,000 [95% CI, 135.34–193.95]) than in European participants (115.78 per 100,000 [95% CI, 102.75–129.99]), while clinical recognition showed the opposite pattern. African ancestry carriers had lower odds of diagnosis than European ancestry carriers (odds ratio, 0.22 [95% CI, 0.09–0.52]; P = 7.16E-04), an association attenuated yet not eliminated by adjustment for healthcare utilization (odds ratio, 0.29 [95% CI, 0.12–0.72]; P = 7.86E-03).

Conclusions and Relevance

Genomically identified SCAs were associated with higher cardiovascular and cardiometabolic disease burden but were rarely clinically recognized, with evidence of ancestry-patterned diagnostic gaps.

Abbreviations

- **BAF:** B-Allele frequency
- **CDR:** Curated data repository
- **CIF:** Cumulative incidence function
- **CV:** Cardiovascular
- **CVD:** Cardiovascular disease
- **EHR:** Electronic health records
- **GDA:** Global diversity array
- **HR:** Hazard ratio
- **ICD:** International classification of diseases
- **LRR:** Log-R ratios
- **OR:** Odds ratio
- **OMOP:** Observational medical outcomes partnership
- **PAF:** Population attributable fraction
- **PH:** Proportional hazard(s)
- **SCA:** Sex chromosome aneuploidy/aneuploidies
- **SCT:** Sex chromosome trisomy / trisomies
- **SNP:** Single nucleotide polymorphism
- **VTE:** Venous thromboembolism
- **WGS:** Whole genome sequencing
- **XOI:** X-chromosome inactivation

Contents