Rare coding variant burden and risk of late effects among childhood cancer survivors in the St. Jude Lifetime Cohort Study (SJLIFE) and the Childhood Cancer Survivor Study (CCSS)

SJLIFE Working groups: Genomics/Genetics
CCSS Working Groups: Genetics; SMN; Chronic Health Conditions

Investigators

Chang Li	chang.li@stjude.org
Achal Neupane	achal.neupane@stjude.org
Kendrick Li	kendrick.li@stjude.org
Yutaka Yasui	yutaka.yasui@stjude.org
Kevin Krull	kevin.krull@stjude.org
Monica Gramatges	gramatge@bcm.edu
Eric Chow	ericchow@uw.edu
Kevin Oeffinger	kevin.oeffinger@duke.edu
Joe Neglia	pedchair@umn.edu
Lucie Turcotte	turc0023@umn.edu
Stephanie Dixon	stephanie.dixon@stjude.org
Dan Mulrooney	Daniel.mulrooney@stjude.org
Matt Ehrhardt	matt.ehrhardt@stjude.org
Deokumar Srivastava	kumar.srivastava@stjude.org
Kirsten K. Ness	kiri.ness@stjude.org
Melissa Hudson	melissa.hudson@stjude.org
Gregory T. Armstrong	greg.armstrong@stjude.org
Yadav Sapkota	yadav.sapkota@stjude.org

Background and Rationale

Advances in childhood cancer treatment over recent decades have significantly increased survival rates, with more than 85% of diagnosed individuals living at least five years beyond their initial diagnosis[1–3], leading to a substantial population of long-term survivors. However, this success comes with a substantial burden of chronic health conditions that may emerge years or decades after treatment. These late effects span a wide range of disease domains, including cardiovascular, endocrine, neurological, pulmonary, and second malignant neoplasms that significantly affect survivors' quality of life and long-term morbidity and mortality risk[4,5].

Previous research has demonstrated that cancer treatment exposures, such as radiation therapy and chemotherapy, are strongly associated with the development of these late effects[6,7]. Nonetheless, substantial inter-individual variation in risk remains even among survivors with similar treatment histories, suggesting that inherited genetic susceptibility plays a critical role[4]. While genome-wide association studies (GWASs) have identified several common variants linked to late-effect phenotypes in survivors[8–12], the contribution of rare variants remains largely unexplored. In the general population, rare coding variants, particularly those that are deleterious or loss-of-function, have been shown to exert larger effects and exhibit higher penetrance than common variants[15,16]. These rare variants are increasingly implicated in complex traits and diseases across large-scale sequencing efforts in the general population, such as the UK Biobank[17] and Icelandic[18] population studies. Importantly, advanced statistical methods, including the aggregation of rare variants across genes through burden testing or kernel-based methods and expanded polygenic scores (EPGS) have been shown to substantially improve statistical power in identifying gene-level associations[15,19,20]. Despite these advances, the role of rare coding variants in modifying the risk of late effects in childhood cancer survivors has not been systematically investigated. This gap represents a critical opportunity to expand our understanding of gene-environment interactions in a population exposed to intensive therapeutic regimens early in life. The availability of whole-exome sequencing (WES) and whole-genome sequencing (WGS) data, combined with late-effect phenotypes from two well-characterized childhood cancer survivor cohorts, the St. Jude Lifetime Cohort Study

(SJLIFE) and the Childhood Cancer Survivor Study (CCSS), provides a unique opportunity to investigate this question in depth.

Recognizing this critical knowledge gap, we propose a comprehensive investigation into how rare coding variants influence the risk of cardiovascular disease, subsequent neoplasms, and all-cause mortality among childhood cancer survivors, as well as their impact on specific diseases within these categories. Utilizing high-quality whole-genome and whole-exome sequencing data from over 12,000 participants across the well-characterized SJLIFE and CCSS cohorts, we will employ a multi-faceted approach. This involves grouping rare coding variants based on clinical pathogenicity and predicted functional impact, followed by rigorous association testing for both binary and continuous outcomes (only in SJLIFE). It is important to note that this initial phase of analysis will serve as hypothesis-generating, utilizing foundational models and controlling for general covariates, and ignoring highly phenotype-specific covariate adjustments for individual diseases, to identify initial signals of association. We will primarily group late effects by organ system for broad genetic susceptibility exploration. Once early screening results are ready, we will bring relevant SJLIFE clinical experts and clinician scientists into the project to provide domain-specific guidance on data interpretation and ensure appropriate clinical input for subsequent analyses. If sample size permits, we expect to expand our analytical framework to those individual phenotypes. Furthermore, we will perform time-to-event analyses to investigate the temporal impact of these rare variants on the time of onset for incident late effects. Finally, to enhance personalized risk stratification, we will also develop and validate an EPGS, integrating both common and rare variant effects. All analyses will be conducted independently within each cohort, with subsequent meta-analysis for comparable phenotypes shared across datasets to enhance statistical power and generalizability.

Specific Aims

Aim 1: To identify gene-level rare coding variant burdens associated with the prevalence or overall occurrence of late-effect phenotypes, including mortality, in childhood cancer survivors.

- Our study will preliminarily focus on three main composite late effects: cardiovascular disease, subsequent neoplasm, and all-cause mortality. Individual diseases or cause or mortality with a minimum of 30 cases will also be included. For less prevalent diseases, inclusion will be determined by the performance observed in the major late effects mentioned previously, and active collaboration with relevant working groups will be pursued. When sufficient sample sizes are available, parallel analyses will be conducted in both the SJLIFE and CCSS cohorts for binary late effects. The results from these separate analyses will then be meta-analyzed to maximize the power for identifying a genome-wide signal.

  Aim 1a: To identify gene-level rare coding variant burden associated with selected late-effect phenotypes and mortality (all-cause and cause-specific) in childhood cancer survivors.

  - This will establish initial associations between variant burden and both the presence of chronic health conditions and the occurrence of mortality events.

  Aim 1b: To examine associations stratified by CTCAE grade and treatment exposures.

  - For phenotypes graded using the Common Terminology Criteria for Adverse Events (CTCAE) v4.03 grading system[4], we will stratify analyses comparing CTCAE 0 grade to those >=2 and to those >=3, to assess whether rare variant burden is associated with severe phenotypes or mortality (e.g., CTCAE grade 5). For a composite phenotype, the highest CTCAE grading across all observed events for a survivor will be used. A stratified analysis will be conducted for survivors who received radiotherapy and/or chemotherapy, particularly anthracycline.

  Aim 1c: To characterize variant-level contributions within significant genes.

- We will examine top signals from gene-level tests to assess effect size distributions, individual variant contributions, carrier status among cases and controls.

Aim 2: To comprehensively characterize the temporal impact of rare coding variant burden on the age of onset for incident late-effect phenotypes and the age at death for mortality (all-cause and cause-specific) in childhood cancer survivors.

- We plan to perform Cox proportional hazards models for all candidate genes (or gene-level rare variant burdens) to assess their associations with the age of onset for incident late effects and the age at death for mortality. The analysis will then be stratified by treatment exposure. This aim will provide a dynamic, longitudinal understanding of genetic risk, discerning whether identified rare variant burdens accelerate or delay the onset of chronic health conditions or contribute to earlier mortality. This is crucial for understanding how genetic factors affect the overall survivorship trajectory, including the ultimate outcome of death.

Aim 3: To develop and validate an expanded polygenic score integrating common and rare variant effects for improved risk stratification of late effects and all-cause mortality.

- Building on the gene-level rare variant signals from Aim 1, this aim will construct and validate an expanded polygenic score (EPGS). We then aim to determine if this combined score significantly improves the identification of childhood cancer survivors at high risk for late effects including mortality compared to a PGS based on common variants alone. This will allow for risk stratification for both chronic conditions and the ultimate risk of premature death.

Methods

Study Population

We will include all 5-year survivors of childhood cancer from SJLIFE (N=4,467), CCSS expansion (N= 2,837) and CCSS original (N=5,013) cohorts with WES data available. These numbers account for the exclusion of survivors who participated in both SJLIFE and CCSS. All analyses will be conducted independently in each cohort to account for differential treatment eras, cohort characteristics, and technical variability introduced by the use of different WES pipelines across CCSS original and expansion cohorts. Additionally, analyses will also be performed within subgroups stratified by sex, cancer treatment exposures, and genetic ancestry. Additionally, for SJLIFE and CCSS expansion cohorts, where WGS data are available, complementary analyses leveraging WGS data will be conducted where appropriate, e.g. calculation of common variant polygenic risk scores in Aim 3.

Outcome variables

Preliminary analysis is designed as an exploratory analysis of late effect categories in cardiovascular disease, subsequent neoplasm, and all-cause mortality, as well as their individual diseases with sufficient number of events (number of cases >= 30) observed in SJLIFE and CCSS cohorts. For continuous traits related to these late effects from SJLIFE, we will exclude those with greater than 50% missing data.

In SJLIFE, some survivors may have missing medication information, which could influence the accuracy of laboratory-based biomarkers or clinical measures. To address this, we will apply CTCAE-based adjustments, considering a CTCAE grade ≥ 3 (e.g., cardiomyopathy grade ≥ 3) as an indicator that the individual is likely receiving relevant medications, and we will include medication status as a covariate in subsequent models. Because continuous laboratory measures (e.g., ejection fraction, lipids) are available only within SJLIFE, we will adopt a two-stage analytical approach: first, evaluating associations between laboratory results and corresponding clinical outcomes within SJLIFE (e.g., ejection fraction and cardiomyopathy), and then examining related outcomes such as heart failure in the CCSS cohort.

When proven effective, our method will also be applied and expanded to a wide range of late-effect phenotypes, analogous in structure to a phenome-wide association study (PheWAS)[21]. As such, our ultimate goal is to include all available outcome measures relevant to long-term survivorship in childhood cancer, leveraging the deep phenotyping available in both SJLIFE and CCSS. In SJLIFE, we will analyze both binary and continuous outcomes, taking full advantage of the longitudinal clinical assessments, laboratory measures, and quantitative trait data collected in this cohort. In the CCSS cohorts (both original and expansion), analyses will be limited to binary outcomes, as continuous phenotype data are not available.

- Chronic health conditions, as graded by the Common Terminology Criteria for Adverse Events (CTCAE) v4.03 grading system[4]. Analyses will consider comparisons between 0 with >=2 (moderate and worse), and 0 with >=3 (severe or life-threatening and worse) as appropriate for the late effect.
- Continuous traits related to our primary late effect of interest that are available in SJLIFE, such as laboratory biomarkers (e.g., lipid levels), anthropometric measurements (e.g., BMI, blood pressure), and echocardiogram/ electrocardiogram measurements.
- Mortality outcomes will include all-cause mortality and cause-specific mortality from cardiac events and subsequent neoplasms.

Sociodemographic/clinical variables

- Age at measurements of continuous traits
- Age at last contact (prevalent analysis)
- Age at blood sample collection (incident analysis)
- Sex
- Age at primary cancer diagnosis
- Cancer treatment exposures within 5 years of primary cancer diagnosis
    - Any RT (yes/no)
    - Field-specific RT with total body irradiation (TBI) (yes/no)
    - Any chemotherapy (yes/no)
    - Any surgery (yes/no)

Genetic data

For SJLIFE survivors and CCSS survivors, we will use quality-controlled WES data. Additionally, WGS data will be available and utilized for SJLIFE and CCSS expansion cohorts when calculating polygenic scores in Aim 3. In both SJLIFE and CCSS, genetic ancestry (European, African and East Asian) of survivors will be determined using a K-means clustering approach implemented in Admixture[22], on the basis of genotype data of an independent set of common autosomal SNVs and the 1000 Genomes Project samples as ancestral populations. Survivors will then be grouped into European (%European >80%), African (%African >60%) and Others based on the estimated ancestry proportions.

P/LP variants

Rare variants will be grouped into genes using four functionally distinct masks to define potential pathogenicity in a nested manner where the last mask will capture all missense variants. The first mask includes variants classified as pathogenic or likely pathogenic in ClinVar[26], consistent with previous studies[27,28], and will be limited to entries without conflicting interpretations that were curated by clinical laboratories from 2015 onward. The second mask consists of predicted high-confidence loss-of-function (pLoF) variants identified using LOFTEE[25], which filters for frameshift indels, stop-gain variants, and splice-disrupting mutations while excluding variants flagged as low confidence. The third mask includes predicted deleterious missense variants, annotated using SnpEff[23] and classified based on a consensus of over 90% agreement across in silico prediction tools within dbNSFP[24] (version 5.1a). The last mask will be all missense variants. The rationale for these masks are described in **Statistical analysis Aim 3**.

Statistical analysis

Aim 1: To identify gene-level rare coding variant burdens associated with the prevalence or overall occurrence of late-effect phenotypes, including mortality, in childhood cancer survivors.

To evaluate the association between rare coding variants and late-effect phenotypes, i.e. cardiovascular diseases, subsequent neoplasms and mortality (all-cause and cause-specific), we will conduct gene-based burden testing independently in the SJLIFE, CCSS original, and CCSS expansion cohorts (accounting for under-sampling of acute lymphoblastic leukemia survivors). For each (composite) phenotype, we will use gene-level aggregation of rare variants, applying appropriate regression models based on the outcome type. For binary phenotypes, we will perform gene-based rare variant burden testing using two complementary statistical frameworks that address distinct analytical challenges commonly encountered in rare variant analyses. First, we will employ Firth bias-corrected logistic regression for burden-based test, implemented in the logistf R package. This method is particularly advantageous for rare variant analyses due to its ability to mitigate small-sample bias and address issues like separation or non-convergence, thereby providing robust initial estimates of the overall odds of disease associated with rare variant burden, even when dealing with low-frequency variant carriers or sparse outcome distributions. At the same time, a kernel based SKAT-O[19] test using the SKAT R package will be performed to account for the possible heterogeneous effects of variants. Second, we will apply REGENIE[30] (v4.1), a two-step whole-genome regression framework that efficiently handles large-scale genotype data while accounting for relatedness and population structure. REGENIE is well-suited for rare variant testing in binary outcomes, especially when case-control imbalance is present, due to its ability to leverage ridge regression in the first stage to estimate polygenic background and phenotype prediction, followed by fast and flexible association testing. We will implement REGENIE using the leave-one-chromosome-out (LOCO) approach to avoid proximal contamination during model training and ensure unbiased effect size estimation. For continuous outcomes, available only in the SJLIFE cohort, we will adopt a similar dual-analytic strategy. Standard linear regression models will be used to estimate associations between gene-based rare variant burden and quantitative traits, adjusting for relevant covariates such as age at diagnosis, sex, ancestry, and treatment exposures. In parallel, we will apply REGENIE's linear regression module to the same set of outcomes and covariates.

For phenotypes measured in more than one cohort, we will conduct inverse-variance weighted meta-analysis across SJLIFE, CCSS original, and CCSS expansion cohorts. Meta-analysis will be performed using metal[31], applying fixed-effect models under homogeneity and random-effects models when heterogeneity is detected (assessed via Cochran's Q[32] test and I Index[33]). With only three studies, the pooled estimates will likely have limited stability. The small number makes it difficult to accurately estimate heterogeneity and increases the influence of individual studies. Confidence intervals will be wide, and any conclusions will be interpreted cautiously, especially for the random-effect models. Variants exhibiting evidence of ancestral allelic heterogeneity (Phet <0.05) will be meta-analyzed using the Han-Eskin random-effects model (RE2) in METASOFT[34]. Where possible, ancestry-stratified meta-analysis will also be conducted to explore population-specific effects and increase the generalizability of findings.

To better understand the relationship between variant burden and late-effect CTCAE grading, we will stratify analyses for CTCAE-graded phenotypes. This will involve separately examining associations for outcomes of grade ≥2, grade ≥3, and grade=5 (death) against unaffected individuals. Analyzing grade ≥ 2 and grade ≥ 3 separately allows us to capture differences in severity, clinical impact, and underlying risk factors without diluting signals by pooling heterogeneous outcomes. For significant signals, we will further investigate the distribution of variant carriers among survivor cases and controls, explore heterogeneity in effect sizes, and report individual variant-level metrics. These metrics will include carrier counts, odds ratios, and allele frequencies from external reference datasets like gnomAD[35]. Crucially, when exploring effect size heterogeneity, our primary focus will be on stratifying analyses based on major cancer treatments received, such as radiotherapy and chemotherapy. This approach will help us understand how treatment exposures interact with rare genetic variants to influence the risk of late effects. Although this may reduce sample size in individual tests, the potential for increased homogeneity could improve statistical power for certain survivor subgroups. We will use visualization tools such as forest plots and lollipop plots to highlight key associations.

Aim 2: To comprehensively characterize the temporal impact of rare coding variant burden on the age of onset for incident late-effect phenotypes and the age at death for mortality in childhood cancer survivors.

To fully understand how rare coding variants temporarily affect incident late-effect phenotypes, we will use Cox proportional hazards models for all gene-level rare variant burdens in our cohorts. This method directly models the immediate risk of late effect onset over time or mortality, helping us determine if specific rare variant burden speeds up or slows down the time of onset for incident late effects. These models will assess the relationship between the gene-level rare variant burden (from Aim 1a) and the time until specific late effects begin, with the time scale starting five years after the primary cancer diagnosis (or time of DNA sample collection for mortality analyses). All models will be adjusted for the same covariates used in Aim 1, including age at primary cancer diagnosis, various cancer treatment exposures, sex, and PCs for genetic ancestry. We will specifically check the proportional hazards assumption for all covariates. If violations occur for key genetic factors, we will use appropriate extensions, such as time-varying coefficients, to model the dynamic nature of genetic risk. Additionally, similar to Aim 1, a stratified analysis will be performed for survivors who received radiotherapy and/or chemotherapy. This will help us investigate how these treatment exposures influence the temporal impact of rare genetic variants on the age of onset for incident late effects and age at death.

Aim 3: To develop and validate an expanded polygenic score integrating common and rare variant effects for improved risk stratification of selected late effects and all-cause mortality.

To develop and validate an expanded polygenic score (EPGS) that integrates common and rare variant effects for improved risk stratification of selected late effects and all-cause mortality, we will construct two score components and then calibrate their combined contribution to outcome risk within the target cohort. The common-variant PGS for each late-effect phenotype will be selected in collaboration with clinical investigators. When available, we will adopt existing scores from the PGS Catalog that match the target ancestral population and phenotype; otherwise, we will derive scores from suitable GWAS summary statistics matched by ancestry or from existing SJLIFE publications. For the rare-variant PGS, we will adapt the nested variant-weighting framework introduced by Dornbos et al.[36], tailoring it to our four-mask design to address the challenge of weighting ultra-rare variants whose individual effects cannot be estimated reliably.

The rare-variant framework proceeds in two steps: gene selection and variant weighting. In step one, we will employ a “loose” gene selection criterion to maximize sensitivity while preserving biological plausibility. Genes will be included if they show nominal association with the late effect in Aim 1 burden testing (for example, $P < 0.05$) and/or have independent evidence of relevance to the phenotype through curated pathway membership or orthogonal biological data. Pathway evidence will be curated from domains directly pertinent to survivorship biology and treatment toxicities, for example the DNA damage repair pathway (GO:0006281).

In step two, we will implement nested mask-based weighting within each selected gene using our four functional masks (ClinVar Pathogenic/Likely Pathogenic curated since 2015 without conflicting interpretations; high-confidence pLoF by LOFTEE; deleterious missense by dbNSFP v5.1a with ≥90% consensus; and all other missense variants). For each gene-by-mask combination, we will estimate an aggregate effect size in the discovery cohort(s) using gene-level burden models appropriate to the outcome (e.g., logistic regression for binary endpoints, Cox models for time-to-event), evaluated across prespecified rare-frequency thresholds (e.g., MAF < 1% and < 0.1%). Masks are ordered from most stringent to most inclusive, and each variant inherits the weight from the most stringent mask it qualifies for. This nested assignment prevents double-counting, assigns larger weights to more deleterious variants, and provides principled weights to variants not observed in discovery as long as they map to a mask. An individual’s rare-variant PGS is computed as the sum of these assigned weights across all qualifying variants they carry in the included genes, accounting for allele dosage where applicable.

After computing the common and rare components, we will perform an explicit calibration step to map scores to risk in the target cohort, while keeping per-variant and per-gene weights fixed. For binary outcomes, we will fit a logistic regression model of the form

$$logitP(Y = 1) = \alpha + \gamma_c \tilde{S}\text{common} + \gamma_r \tilde{S}\text{rare} + X$$

where $\tilde{S}_{\text{common}}$ and $\tilde{S}_{\text{rare}}$ are z-scored within the training split and $\mathbf{X}$ comprises covariates. This "light" training calibrates the scaling and intercept to the target cohort without re-estimating individual SNP or variant weights, thereby preserving comparability and minimizing overfitting. Score standardization and calibration parameters will be estimated exclusively in the training split and then applied to held-out data. Alternatively, we can use SJLIFE as our training/exploration set, and using CCSS as test/replication set.

To ensure a fair comparison between the EPGS and the common PGS alone, we will apply an identical calibration protocol, covariate set, and data splits to both models. Specifically, we will fit a calibrated baseline model with only the common PGS and covariates, and a calibrated EPGS model that adds the rare component, using the same training data and regularization settings. We will then evaluate both models on a strictly held-out test set. For time-to-event outcomes, we will compare Harrell's C-index[37] and summarize time-dependent AUC[38] across prespecified follow-up windows, alongside calibration diagnostics and the integrated Brier score. For binary outcomes, we will assess discrimination (AUROC, and AUPRC in imbalanced settings), calibration-in-the-large and slope, Brier score, and decision-curve analysis; intercept-only recalibration will be performed if outcome prevalence differs substantially from the training split. We will quantify incremental value using nested model comparisons (e.g., likelihood ratio tests), net reclassification improvement, and integrated discrimination improvement, and we will report uncertainty via bootstrap or cross-validation where appropriate.

All analyses will be conducted with careful attention to ancestry and transportability. Where available, we will use ancestry-matched weights for the common PGS; otherwise, we will use multi-ancestry or best-available estimates and include genetic principal components to adjust for population structure. Prior to combining components, we will perform conditional and LD-aware sensitivity analyses to verify that common- and rare-variant signals are largely independent. Model development, calibration, and evaluation will adhere to a prespecified analysis plan with fixed splits to prevent information leakage, and all final performance metrics will be reported on the held-out test set to provide an unbiased estimate of clinical utility.

Impact statement

While common genetic variants explain only a fraction of late-effect risk in childhood cancer survivors, the contribution of rare variants remains largely unexplored. This project will conduct the first large-scale, gene-level analysis of rare coding variants across a wide range of late effects, providing a foundational resource to generate new hypotheses about the biological mechanisms of treatment-related toxicity.

Example tables and figures

Table 1. Characteristics of childhood cancer survivors from SJLIFE and CCSS.

Characteristics	SJLIFE	CCSS original	CCSS expansion
Age at primary cancer diagnosis (years)			
Age at last contact (years)			
Sex			
Male			
Female			
Radiation Therapy			
Any radiation			
Total body irradiation			
Chemotherapy			
Any			
Surgery			
Any			

Figure 1. Overview of study design and analysis workflow

Figure 2. Distribution of rare variant carrier burden by phenotype group

- A boxplot or violin plot showing the distribution of rare variant carrier burden per gene across participants, stratified by major phenotype groups (e.g., cardiovascular, endocrine, neurocognitive, second malignancy). Carrier burden is based on the number of qualifying rare variants (per mask) per individual. Separate panels represent the three variant masks (deleterious missense, LOF, ClinVar P/LP). SJLIFE is shown as the primary cohort; binary traits from CCSS cohorts may be overlaid or presented in supplementary figures.

Figure 3. Gene-level associations for late effects

- Manhattan style plot or volcano plot displaying gene-level association results for each late effect. The x-axis shows genes, and the y-axis represents log10 (p-value) from burden tests using the P/LP masks and MAF <0.1%. Highlighted genes pass multiple testing correction thresholds. Labels indicate gene names with strongest associations. Sensitivity analyses by ancestry and treatment subgroup will be shown using colored points or confidence intervals.

Figure 4. Association of rare variant burden with age of onset for incident late effects.

- Kaplan-Meier curves or forest plots derived from Cox proportional hazards models will be used to show how different levels of rare variant burden (e.g., presence vs. absence of burden in a significant gene, or high vs. low EPGS quintiles) are associated with the cumulative incidence or age of onset for specific incident late effects. Separate panels could show results for different significant genes or EPGS strata.

Figure 5. Predictive performance of the Expanded Polygenic Score (EPGS) for late-effect risk.

- Receiver operating characteristic (ROC) curves with corresponding Area Under the Curve (AUC) values and confidence intervals for binary outcomes, and potentially R-squared values for continuous outcomes. Separate panels could show results for different late-effect categories (e.g., cardiovascular disease, metabolic syndrome). The aim is to visually demonstrate the incremental predictive value gained by incorporating rare variant burden.

References

1. Armstrong, G. T. Reduction in Late Mortality among 5-Year Survivors of Childhood Cancer. *New Engl J Med* **374**, 833–842 (2016).

2. Winther, J. F. Childhood cancer survivor cohorts in Europe. *Acta Oncologica* **54**, 655–668 (2015).

3. Fidler, M. M. Long term cause specific mortality among 34489 five year survivors of childhood cancer in Great Britain: population based cohort study. *Bmj-Brit Med J* **354**, (2016).

4. Hudson, M. M. Approach for Classification and Severity Grading of Long-term and Late-Onset Health Events among Childhood Cancer Survivors in the St. *Jude Lifetime Cohort. Cancer Epidemiol Biomarkers Prev* **26**, 666–674 (2017).

5. Robison, L. L. & Hudson, M. M. Survivors of childhood and adolescent cancer: life-long risks and responsibilities. *Nat Rev Cancer* **14**, 61–70 (2014).

6. Armstrong, G. T., Stovall, M. & Robison, L. L. Long-Term Effects of Radiation Exposure among Adult

Survivors of Childhood Cancer: Results from the Childhood Cancer Survivor Study. *Radiation research* **174**, 840–850 (2010).

7. Neupane, A. Contributions of cancer treatment and genetic predisposition to risk of subsequent neoplasms in long-term survivors of childhood cancer: a report from the St Jude Lifetime Cohort and the Childhood Cancer Survivor Study. *The Lancet. Oncology* **26**, 806–816 (2025).

8. Dong, Q. Genome-wide association studies identify novel genetic loci for epigenetic age acceleration among survivors of childhood cancer. *Genome Medicine* **14**, (2022).

9. Sapkota, Y. Genome-Wide Association Study in Irradiated Childhood Cancer Survivors Identifies HTR2A for Subsequent Basal Cell Carcinoma. *J Invest Dermatol* **139**, 2042- (2019).

10. Sapkota, Y. Whole-Genome Sequencing of Childhood Cancer Survivors Treated with Cranial Radiation Therapy Identifies 5p15.33 Locus for Stroke: A Report from the St. *Jude Lifetime Cohort Study. Clin Cancer Res* **25**, 6700–6708 (2019).

11. Richard, M. A. Germline Genetic and Treatment-Related Risk Factors for Diabetes Mellitus in Survivors of Childhood Cancer: A Report From the Childhood Cancer Survivor Study and St Jude Lifetime Cohorts. *Jco Precis Oncol* **6**, (2022).

12. Im, C. Trans-Ancestral Genetic Risk Factors for Treatment-Related Type 2 Diabetes Mellitus in Survivors of Childhood Cancer. *Journal of Clinical Oncology* **42**, (2024).

13. Im, C. Generalizability of "GWAS Hits" in Clinical Populations: Lessons from Childhood Cancer Survivors. *American Journal of Human Genetics* **107**, 636–653 (2020).

14. Manolio, T. A. Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).

15. Backman, J. D. Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature* **599**, 628- (2021).

16. Forrest, I. S. *et al.* Population-Based Penetrance of Deleterious Clinical Variants. *JAMA* **327**, 350–359 (2022).

17. Weiner, D. J. Polygenic architecture of rare coding variation across 394,783 exomes. *Nature* **614**, 492- (2023).

18. Jensson, B. O. Actionable Genotypes and Their Association with Life Span in Iceland. *New Engl J Med* **389**, 1741–1752 (2023).

19. Lee, S., Wu, M. C. & Lin, X. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* **13**, 762–775 (2012).

20. Williams, J. *et al.* Integrating Common and Rare Variants Improves Polygenic Risk Prediction Across Diverse Populations. (2024) doi:10.1101/2024.11.05.24316779.

21. Bastarache, L., Denny, J. C. & Roden, D. M. Phenome-Wide Association Studies. *JAMA* **327**, 75–76 (2022).

22. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19**, 1655–1664 (2009).

23. Cingolani, P. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w(1118); iso-2; iso-3. *Fly* **6**, 80–92 (2012).

24. Liu, X., Li, C., Mou, C., Dong, Y. & Tu, Y. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Genome Med* **12**, 103 (2020).

25. Karczewski, K. J. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).

26. Landrum, M. J. ClinVar: improvements to accessing data. *Nucleic acids research* **48**, 835–844 (2020).

27. Ke, H. Landscape of pathogenic mutations in premature ovarian insufficiency. *Nature Medicine* **29**, 483–492 (2023).

28. Shekari, S. *et al.* Penetrance of pathogenic genetic variants associated with premature ovarian insufficiency. *Nat Med* **29**, 1692–1699 (2023).

29. Ganna, A. *et al.* Quantifying the Impact of Rare and Ultra-rare Coding Variation across the Phenotypic Spectrum. *Am J Hum Genet* **102**, 1204–1211 (2018).

30. Mbatchou, J. Computationally efficient whole-genome regression for quantitative and binary traits. *Nat Genet* **53**, 1097- (2021).

31. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).

32. Cochran, W. G. The Combination of Estimates from Different Experiments. *Biometrics* **10**, 101–129 (1954).

33. Ioannidis, J. P. A., Patsopoulos, N. A. & Evangelou, E. Heterogeneity in Meta-Analyses of Genome-Wide Association Investigations. *PLoS One* **2**, (2007).

34. Han, B. & Eskin, E. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *Am J Hum Genet* **88**, 586–598 (2011).

35. Chen, S. A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* **625**, 92–100 (2024).

36. P, D. *et al.* A combined polygenic score of 21,293 rare and 22 common variants improves diabetes diagnosis based on hemoglobin A1C levels. *PubMed*.

37. Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B. & Wei, L. J. On the C-statistics for Evaluating Overall Adequacy of Risk Prediction Procedures with Censored Survival Data. *Stat Med* **30**, 1105–1117 (2011).

38. Kamarudin, A. N., Cox, T. & Kolamunnage-Dona, R. Time-dependent ROC curve analysis in medical research: current methods and applications. *BMC Medical Research Methodology* **17**, 53 (2017).

St. Jude - Confidential