

# Predictors of early-onset cancer risk: insights from machine learning analyses of the Christchurch Health and Development Study data

Simranjeet S Dahia, Laalithya Konduru, Joseph M Boden, Savio G Barreto

The global incidence of early-onset cancers in adults is rising, posing a significant public health challenge.<sup>1</sup> Unlike cancers in older populations, the risk factors for early-onset cancers remain poorly understood. Perinatal and early life stressors have been postulated to modulate the risk of early-onset adult cancers.<sup>2</sup> A prospective lifecourse cohort study provides the ideal framework to study perinatal and early-life stressors. Existing prospective lifecourse cohorts were established to study childhood diseases and chronic health conditions to test the developmental origins of health and disease (DOHaD) hypothesis.<sup>3</sup> These cohorts, thus, lack cancer-specific variables and have low (cancer) event rates, reducing statistical power.

Traditional epidemiological methods, such as univariate screening with false discovery rate (FDR) correction (a statistical method used to control the rate of false positives when performing multiple hypothesis tests), frequently yield null results in high-dimensional datasets with rare outcomes due to low power and inability to capture complex interactions.<sup>4</sup> Some studies using such datasets and methods failed to identify significant associations between perinatal factors and early-onset cancers due to these limitations.<sup>5</sup> To overcome these challenges, we applied a hybrid machine learning (ML) and logistic regression pipeline to data from the Christchurch Health and Development Study (CHDS) to identify perinatal predictors of early-onset cancers in adults, with an aim to gain insight into the underlying aetiological pathways for these cancers.

## Methods

We analysed data of the cohort members of the CHDS—a longitudinal birth cohort study of children born in the Christchurch, New Zealand urban region over a 4-month period during 1977.<sup>6</sup>

We applied a hybrid pipeline<sup>7</sup> that integrated ML feature selection (LASSO and tree-based models) with multivariable logistic regression to estimate odds ratios (ORs) to overcome the challenges with univariate testing (with FDR correction) inherent to a wide dataset. Analyses were conducted in Python (Python algorithm version 3.12.5, Python Software Foundation, Wilmington, DE, USA), leveraging a high-performance computing cluster for computational power.

## Results

Among the 1,265 children included in the CHDS, there are 41 recorded cancer cases by the age of 40 years within the cohort (target variable), along with 158 perinatal, demographic and lifestyle factors (predictors), resulting in a wide dataset.

The hybrid pipeline identified four significant factors associated with early-onset cancer risk. Antenatal care provided by a hospital or clinic (OR 3.16, 95% CI 1.30–7.68), antenatal vitamin use (OR 3.96, 95% CI 1.40–11.20) and antenatal cough medicine use (OR 2.96, 95% CI 1.07–8.20) were associated with increased odds of cancer by age 40 years. Conversely, antenatal care by a family doctor (OR 0.28, 95% CI 0.10–0.79) was associated with reduced odds. The ML model achieved an area under the receiver operating characteristic curve of 0.78 (95% CI 0.72–0.84) and a Brier score of 0.12, indicating robust predictive performance.

## Discussion

Our study adds to the literature by identifying novel predictors using a hybrid ML approach, overcoming the low power and high-dimensionality challenges. By detecting associations missed by traditional methods, our findings support the hypothesis that perinatal stressors are predictive of early-onset cancer risk.<sup>2</sup>

The identified factors may modulate early-onset cancer risk through epigenetic processes and metabolic pathways. Patients selected to undergo hospital-based antenatal care are more likely to represent high-risk pregnancies, with attendant maternal stressors (e.g., chronic illness, infections, eclampsia, etc.) that can induce foetal epigenetic changes such as DNA methylation.<sup>8</sup> Conversely, pregnant females deemed suitable to undergo antenatal care with their family doctors would be more likely to have low-risk pregnancies, with consequent lower maternal and foetal stressors. Maternal vitamin supplementation (including folic acid in larger doses) and the risk of cancer in the offspring is a controversial finding that has been previously reported.<sup>9</sup> The increased risk of colon and breast cancers is postulated to occur via disruption of foetal metabolic pathways and modulation of DNA methylation.<sup>9</sup> Maternal cough medicine as a predictor of early-onset cancer in the offspring more likely concurs with previous evidence implicating underlying respiratory infections as a risk factor,<sup>10</sup> rather than the actual cough medicine used.

The limitations of the study include the reliance on a single cohort, limiting generalisability. Also, the observational design of the cohort precludes

derivation of causal inference. There was also no adjustment for confounding. The CHDS did not collect information on cancer type; this limits the ability to analyse associations by cancer subtype. The dataset included 158 parameters spanning the spectrum of perinatal, demographic and lifestyle factors, with the main aim of studying the same cohort of children at repeated intervals throughout their lifecourse. While over the course of the study, this resulted in over 40 million characters of data, the data lacked granularity for specific matters, such as type of vitamin used or if the cough medicine was used for a viral illness where the patient also received paracetamol or another over the counter medication. However, future mediation analyses of the CHDS data will assess whether combined perinatal and early-life stressors increase the risk of early-onset cancer compared to either alone. We will also explore whether ML-identified, non-significant predictors become significant in combination with other variables.

The findings of this study demonstrate novel information on the perinatal predictors of early-onset cancer development relevant to the CHDS cohort, presenting avenues for future research exploring mechanistic pathways.

**COMPETING INTERESTS**

The authors declare no conflict of interest.

**ACKNOWLEDGEMENTS**

This study received funding from Flinders Foundation and Flinders University (Pure ID: 149769468). SGB reports support from the Flinders Foundation Grant: 49358025, the NHMRC Ideas Grant: 2021009, the Pankind 21.R7.INV.CB.UOSA.6.2 supported by funds from the CUREator scheme via Brandon BioCatalyst, The Hospital Research Fund and SALHN Enquiry Grant Round. CHDS is supported by the Health Research Council of New Zealand (programme grant 16/600) and the New Zealand Ministry of Business, Innovation and Employment Strategic Science Fund.

**AUTHOR INFORMATION**

Simranjeet S Dahia: College of Medicine and Public Health, Flinders University, South Australia, Australia.  
Laalithya Konduru, MD: College of Medicine and Public Health, Flinders University, South Australia, Australia.  
Joseph M Boden, PhD: Department of Psychological Medicine, University of Otago, Christchurch, New Zealand.  
Savio G Barreto, FRACS, PhD: College of Medicine and Public Health, Flinders University, South Australia, Australia; Division of Surgery and Perioperative Medicine, Flinders Medical Center, Bedford Park, Adelaide, South Australia, Australia.

**CORRESPONDING AUTHOR**

Savio G Barreto, FRACS, PhD: Department of Surgery, Flinders Medical Centre, Bedford Park, South Australia, Australia 5042. E: georgebarreto@yahoo.com; savio.barreto@sa.gov.au.

**URL**

<https://nzmj.org.nz/journal/vol-138-no-1627/predictors-of-early-onset-cancer-risk-insights-from-machine-learning-analyses-of-the-christchurch-health-and-development-study-d>

**REFERENCES**

- Ledford H. Why are so many young people getting cancer? What the data say. *Nature*. 2024 Mar;627(8003):258-260. doi: 10.1038/d41586-024-00720-6.
- Barreto SG, Pandol SJ. Young-Onset Carcinogenesis - The Potential Impact of Perinatal and Early Life Metabolic Influences on the Epigenome. *Front Oncol*. 2021 Apr 29;11:653289. doi: 10.3389/fonc.2021.653289.
- Baker JL, Gordon-Dseagu VLZ, Voortman T, et al. Lifecourse research in cancer: context, challenges, and opportunities when exploring exposures in early life and cancer risk in adulthood. *Health Open Res*. 2025 Mar 14;6:16. doi: 10.12688/healthopenres.13748.3.
- Savas S, Xu J, Werdyani S, et al. A Survival Association Study of 102 Polymorphisms Previously Associated with Survival Outcomes in Colorectal Cancer. *Biomed Res Int*. 2015;2015:968743. doi: 10.1155/2015/968743.
- Gausman V, Liang PS, O'Connell K, et al. Evaluation of Early-Life Factors and Early-Onset Colorectal Cancer Among Men and Women in the UK Biobank. *Gastroenterology*. 2022 Mar;162(3):981-983.e3. doi: 10.1053/j.gastro.2021.11.023.
- Fergusson DM, Horwood LJ. The Christchurch Health and Development Study: review of findings on child and adolescent mental health. *Aust N Z J Psychiatry*. 2001 Jun;35(3):287-96. doi: 10.1046/j.1440-1614.2001.00902.x.
- Dahia SS, Konduru L, Barreto SG. A Hybrid Machine Learning–Logistic Regression Pipeline for Risk Factor Identification in High-dimensional Epidemiological Data with Extremely Low Events Per Variable. PREPRINT (Version 1) Research Square. 2025. doi: 10.21203/rs.3.rs-7741957/v1.
- Dieckmann L, Czamara D. Epigenetics of prenatal stress in humans: the current research landscape. *Clin Epigenetics*. 2024 Feb 2;16(1):20. doi: 10.1186/s13148-024-01635-9.
- Ciappio ED, Mason JB, Crott JW. Maternal one-carbon nutrient intake and cancer risk in offspring. *Nutr Rev*. 2011 Oct;69(10):561-71. doi: 10.1111/j.1753-4887.2011.00424.x.
- He JR, Hirst JE, Tikellis G, et al. Common maternal infections during pregnancy and childhood leukaemia in the offspring: findings from six international birth cohorts. *Int J Epidemiol*. 2022 Jun 13;51(3):769-777. doi: 10.1093/ije/dyab199. Erratum in: *Int J Epidemiol*. 2022 Jun 13;51(3):1037. doi: 10.1093/ije/dyab228.