Toward Accurate Identification of Fontan and TGA in OMOP CDM: Registry-Anchored Algorithm Validation

Seohu Lee[1], Suhyun Kim[2,3], Haeun Lee[1], Jong M Ko[6], Woo Young Park[4], Kwangsoo Kim[2,3,5], Sang Yun Lee[4], Ari Cedars[6,7]

[1]Biomedical Informatics and Data Science, Johns Hopkins University School of Medicine, Baltimore, MD, USA
[2]Healthcare AI Research Institute and [3]Department of Transdisciplinary Medicine, Seoul National University Hospital, Seoul, Korea
[4]Department of Pediatrics, Seoul National University Children's Hospital, Seoul, Korea
[5]Department of Medicine, College of Medicine, Seoul National University, Seoul, Republic of Korea
[6]Department of Pediatrics and [7]Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA

Background

Adult Congenital Heart Disease (ACHD) encompasses diverse structural heart conditions that persist into adulthood, often requiring surgical correction early in life. Given the heterogeneity and rarity of ACHD, robust multicenter observational research is essential to achieve sufficiently large and diverse cohorts for meaningful analysis[1]. Specifically, conducting large data research involving the complex and rare ACHD subtypes such as Fontan circulation and transposition of the great arteries (TGA) necessitates precise phenotyping strategies[2,3]. The Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) facilitates such efforts by harmonizing data across institutions, supporting systematic comparative analyses and enhancing the generalizability of findings[4].

Prior work using the OMOP CDM has demonstrated the feasibility of grouping physiologically distinct ACHD subtypes via hierarchical code-based algorithms[5]. However, limitations remain, particularly in capturing cases with early childhood surgical repairs and validating definitions against clinical gold standards.

To address these gaps, we aim to refine phenotype definitions for selected ACHD subgroups using multi-source clinical data and evaluate their accuracy across institutions.

Methods

This study utilized longitudinal electronic health record (EHR) data from OMOP CDM instances at Seoul National University Hospital (SNUH) (2004–2024, ~3.8 million patients) and Johns Hopkins Hospital (JHH) (2016–2024, ~2.3 million patients). We defined three surgically significant ACHD physiological subtypes: Fontan circulation, D-TGA with arterial switch, and D-TGA with atrial switch.

Compared to hierarchical algorithms previously developed at Johns Hopkins[5], which broadly categorized ACHD patients into 11 physiologic groups, our refined algorithms focused on these three phenotypes, incorporated explicit exclusion criteria (e.g., Ebstein's anomaly for Fontan), and separated procedure from condition logic to improve specificity. While the prior approach relied on single-institution EHR data, we extended evaluation across two institutions (JHH and SNUH) and integrated additional data sources (Korean medical insurance claims and Korean registry).

For gold standard (GS) comparison, we used a curated billing-based reference set derived from the national Korean billing codes and the Korean Fontan Registry (KFR)[6]. Performance metrics were computed

through patient-level linkage between CDM-extracted cohorts and GS cohorts. True positives (TP) were defined as patients identified by both CDM and GS, false positives (FP) were those identified only by CDM, and false negatives (FN) only by GS. Because D-TGA with arterial switch cases were rare in both claims and registry data, performance evaluation was limited to Fontan and D-TGA with atrial switch.

Result

We first established refined phenotype definitions for the Fontan circulation and D-TGA with arterial or atrial switch operations using OMOP concept IDs (Table 1). The algorithms incorporated both procedure and condition codes, with explicit exclusion criteria to enhance specificity.

Table 1. Fontan and D-TGA Phenotype Definitions

Subtype	Phenotype Criteria	OMOP Concept IDs
Fontan	Procedure occurrences of 'Fontan'	2107268, 2107269, 2107270
	OR	
	Procedure occurrences of 'Glenn'	2107355, 4051948, 2107356, 40491942
	NOT	***NOT***
	Condition occurrences of 'Ebstein's anomaly'	35210812, 4069182
D-TGA (Arterial Switch)	Condition occurrences of 'Discordant ventriculoarterial connection'	432431, 35210794
	AND	
	Procedure occurrences of 'Repair of TGA (anatomic / arterial switch)'	2107375, 2107377
D-TGA (Atrial Switch)	Condition occurrences of 'Discordant ventriculoarterial connection'	432431, 35210794
	AND	
	Procedure occurrences of 'Repair of TGA (non-anatomic / atrial switch)'	2107361, 2107356, 40491942

Application of the algorithms identified 135 Fontan patients at JHH and 297 at SNUH, of whom only a small subset were ≥18 years old (2 and 29, respectively). For D-TGA, 77 arterial switch and 35 atrial switch cases were identified at JHH, while 126 arterial switch and 134 atrial switch cases were found at SNUH. Adult cases were rare, particularly at SNUH where none were ≥18 years.

Table 2. Patient Counts for Fontan and D-TGA Phenotypes at JHH and SNUH

Subtype	JHH (All)	JHH (Age ≥ 18)	SNUH (All)	SNUH (Age ≥ 18)
Fontan	135	2	297	29
D-TGA with Arterial Switch	77	10	126	0
D-TGA with Atrial Switch	35	1	134	0

Algorithm performance was evaluated against billing codes and the KFR as gold standards (Table 3). Compared to claims data, the Fontan CDM phenotype showed precision of 81.8% and recall of 72.8% (F1-score 77.0%). D-TGA with atrial switch demonstrated very high precision (95.5%) but low recall (37.3%), resulting in an F1-score of 53.8%. Compared to registry data, the Fontan CDM phenotype performance was worse (precision 62.6%, recall 44.8%, F1-score 52.3%), as was D-TGA with atrial switch which showed modest precision (62.7%) with low recall (30.4%), yielding an F1-score of 41.0%.

Table 3. Performance Metrics for Fontan and D-TGA (Atrial Switch) in Claims and Registry Data at SNUH

Data Source	Subtype	Count	TP	FP	FN	Precision (%)	Recall (%)	F1-score (%)
Claims	Fontan	334	243	54	91	81.8	72.8	77.0

	D-TGA	341	128	6	215	95.5	37.3	53.8
Registry	Fontan	414	111	229	186	62.6	44.8	52.3
	D-TGA	275	50	192	84	62.7	30.4	41

Conclusion

In this study, we refined ACHD phenotyping algorithms within the OMOP CDM framework, focusing on three surgically significant ACHD subtypes: Fontan circulation, D-TGA with arterial switch, and D-TGA with atrial switch. When applied to longitudinal EHR data from JHH and SNUH, the algorithms identified several hundred patients, although adult cases were relatively rare. This disparity underscores the difficulty of capturing procedures performed in early childhood in adult records using current coding strategies.

Algorithm performance, validated against curated billing-based reference sets and the KFR, demonstrated consistently high precision but variable recall. In claims data, Fontan achieved balanced performance (precision 81.8%, recall 72.8%), whereas D-TGA with atrial switch had very high precision (95.5%) but substantially lower recall (37.3%). Performance compared to registry data was poorer overall, with both phenotypes showing modest precision (≈63%) and limited recall (≤45%). These findings highlight the challenges of ascertaining surgically repaired ACHD cases using structured EHR data alone, particularly for D-TGA, where surgical history is critical but often incompletely captured.

Overall, our results demonstrate the limitations of using CDM-based algorithms alone for multicenter ACHD phenotyping. Future work will focus on integrating multimodal data such as progress note keywords, echocardiography reports and outpatient clinical notes, which often contain diagnostic and surgical context absent from structured fields. At SNUH, LLM-based pipelines are being developed to extract structured information from echocardiography reports as part of the Child Cancer & Rare Disease Project (CCRDP), with planned extensions to narrative clinical data to support a more comprehensive phenotype construction and validation. Finally, broader external validation through organizations such as the Alliance for Adult Research in Congenital Cardiology (AARCC) and international ACHD networks will be essential to enhance the robustness and generalizability of these algorithms.

Acknowledgements

This research was supported and funded by SNUH Lee Kun-hee Child Cancer & Rare Disease Project, Republic of Korea (grant number: 22C-003-0100)

References

1. Verheugt, C. L., Uiterwaal, C. S., van der Velde, E. T., Meijboom, F. J., Pieper, P. G., van Dijk, A. P., ... & Mulder, B. J. (2010). Mortality in adult congenital heart disease. *European heart journal*, *31*(10), 1220-1229.
2. Alsaied, T., Li, R., Christopher, A. B., Fogel, M. A., Slesnick, T. C., Krishnamurthy, R., ... & Rathod, R. H. (2024). High-performing fontan patients: a fontan outcome registry by cardiac magnetic resonance imaging study. *JACC: Advances*, *3*(10), 101254.
3. Shen, H., He, Q., Shao, X., Li, S., & Zhou, Z. (2022). Deep phenotypic analysis for transposition of the great arteries and prognosis implication. *Journal of the American Heart Association*, *11*(3), e023181.
4. Observational Health Data Sciences and Informatics (OHDSI) [Internet]. The Book of OHDSI; [cited 2025 Jun 30]. Available from: http://book.ohdsi.org/
5. Lee S, Ko J, Lee H, Cedars A. Hierarchical Algorithms for Querying Physiologically Distinct Groups in Adult Congenital Heart Disease Using OMOP CDM. *2024 Observational Health Data Sciences and*

Informatics (OHDSI) Global Symposium.

6. Lee SY, Kim SJ, Lee CH, Park CS, Choi ES, Ko H, An HS, Kang IS, Yoon JK, Baek JS, Lee JY, Song J, Lee J, Huh J, Ahn KJ, Jung SY, Cha SG, Kim YH, Lee Y, Cho S. The Long-term Outcomes and Risk Factors of Complications After Fontan Surgery: From the Korean Fontan Registry (KFR). *Korean Circ J*. 2024 Oct;54(10):653-668. https://doi.org/10.4070/kcj.2023.0211