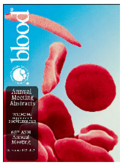
Blood 142 (2023) 7164–7166



The 65th ASH Annual Meeting Abstracts

ONLINE PUBLICATION ONLY

803.EMERGING TOOLS, TECHNIQUES AND ARTIFICIAL INTELLIGENCE IN HEMATOLOGY

A Genotype Validated Bimodal Method for the Large-Scale Identification and Phenotyping of Persons with Sickle Cell Disease Using Electronic Health Record Data

Kristin Wuichet, PhD[1], Clifford M Takemoto, MD[2], Robert Cronin, MDMS[3], Martha Barton, PhD[4], Pei-Lin Chen, MPH[5], Santosh L. Saraf, MD[6], Mitchell J. Weiss, MD PhD[7], Michael R. DeBaun, MD MPH[8]

[1] Vanderbilt University Medical Center, Nashville, TN
[2] Department of Hematology, St. Jude Children's Research Hospital, Memphis, TN
[3] Department of Internal Medicine, The Ohio State University, Columbus, OH
[4] St. Jude Children's Research Hospital, MEMPHIS, TN
[5] St. Jude Children's Research Hospital, Memphis, TN
[6] Department of Medicine, University of Illinois College of Medicine, Willowbrook, IL
[7] Dept. of Hematology, St. Jude Children's Research Hospital, Memphis, TN
[8] Vanderbilt-Meharry Center of Excellence in Sickle Cell Disease, Vanderbilt University Medical Center, Nashville, TN

Introduction

Clinical trials for using hematopoietic stem cell transplant (HSCT) with gene therapies to cure sickle cell disease (SCD) began in 2019. The Food and Drug Administration (FDA) requires that 1) all gene therapy participants receive at least 15 years of ongoing surveillance and 2) their outcomes be compared to a contemporaneous cohort that did not receive gene therapy. Previously we developed an automated contemporaneous cohort of children and adults with SCD from electronic health record (EHR) data (Cronin RM et al. 2023, Blood Adv). Our work was built upon previously explored algorithms utilizing ICD codes and laboratory data to identify persons with sickle cell anemia (SCA), a severe form of SCD (Singh et al. 2018, Blood Adv). Still, our algorithm additionally classifies other predicted genotypes. However, we could not test the SCD phenotypes against the SCD genotypes due to the absence of beta-globin gene sequence data. Here we tested the hypothesis that our Vanderbilt-EHR algorithm for identifying a contemporaneous cohort at Vanderbilt can classify their predicted genotype compared to beta-globin gene sequencing, the gold standard for SCD diagnosis, in the cohort. We subsequently tested the Vanderbilt-EHR algorithm in an established cohort of children from St. Jude Children's Research Hospital that have been deeply phenotyped.

Methods

We applied the Vanderbilt-EHR algorithm to a cohort of 275 samples from Vanderbilt's biorepository of DNA, BioVU, linked to de-identified medical records in a data warehouse called the Synthetic Derivative. These 275 samples were previously submitted for whole genome sequencing as part of the Genetic Variation of Heart, Lung, and Kidney Disease in Sickle Cell Disease: Pre- and Post-Curative Therapies TOPMed project (HLK-SCD phs002617). The returned genotypes were validated using the hemoglobin variant database HbVar (Giardine BM et al. 2021, Nucleic Acids Res) (Table 1). We applied the hemoglobinopathy genotype classification portion of the algorithm to the Sickle Cell Clinical Research and Intervention Program (SCCRIP) cohort from St. Jude Children's Research Hospital, which includes putative genotype diagnoses derived from a comprehensive evaluation of clinical data for each member. We performed statistical analyses of the identification and classification results of the Vanderbilt cohort and the SCCRIP cohort's classification results.

Results

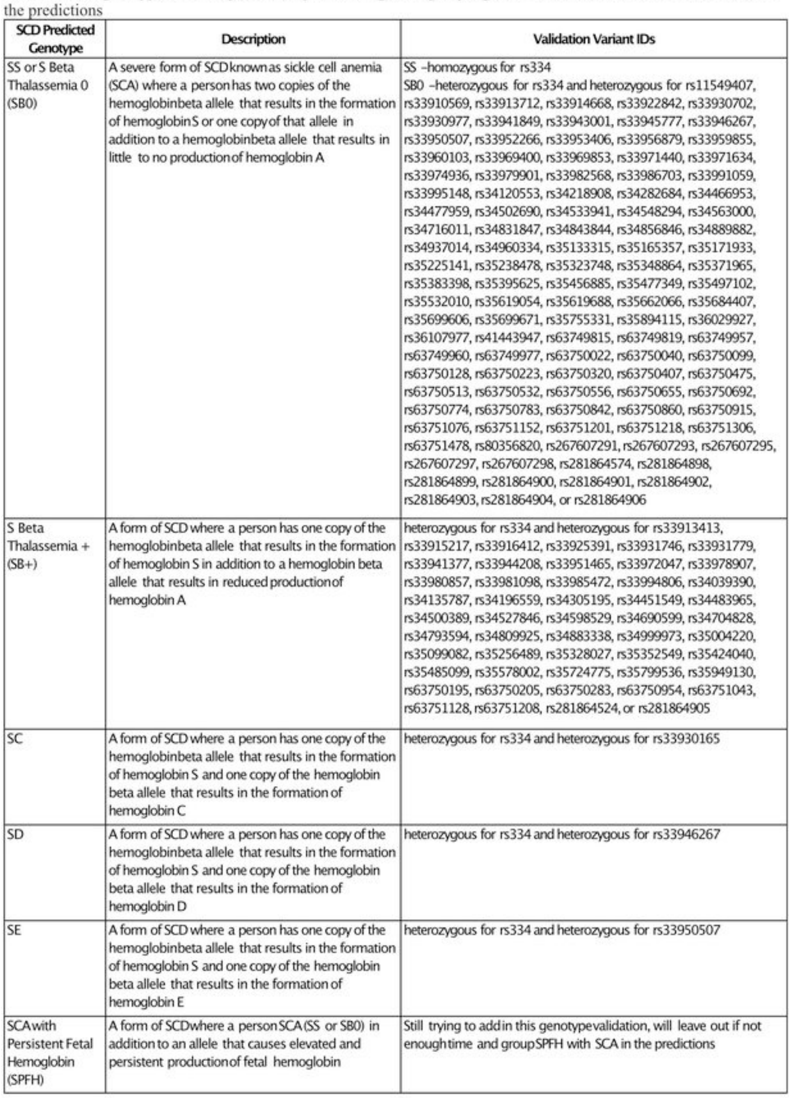
Of the 275 genotyped samples, the algorithm correctly predicted 255 SCD cases and 10 non-SCD cases. There were three cases predicted to be false positives; however, a comprehensive analysis showed that all have laboratory values indicative of SCA (high hemoglobin S levels >60%, absent or negligible hemoglobin A levels, elevated reticulocytes, and low mean corpuscular volume) suggesting that there may be a genotyping error. The remaining 7 cases were either indeterminate in the prediction or in the genotype (Table 1). Given the small number of indeterminate results, we performed a sensitivity analysis to identify the range of performance for SCD identification. The method performed very well with over 98% sensitivity and over 96% positive predictive value (PPV) (Table 2). The lower specificity and negative predictive value (NPV) could be largely attributed to the disproportionately high number of SCD cases in the dataset compared to non-SCD cases. The SCD classification algorithm performed similarly in both the VUMC and SCCRIP cohorts (Table 2). The algorithm performs best in identifying SC and SCA types of SCD with nearly 100% accuracy for SC and over 94% accuracy for SCA. Transfusions can confound the diagnosis of SCA vs. S beta thalassemia +, but the latter has a much lower prevalence consistent with our results (Table 2).

Discussion and Conclusions

Larger SCD cohorts will be identified using these approaches in singular EHR databases and in de-identified data warehouses that EHR companies have developed by aggregating data across multiple EHR systems (e.g., EPIC Cosmos and Cerner Real-World Data). The ongoing need for contemporaneous comparison cohorts of persons with SCD to understand outcomes related to treatment and phenotype can be greatly assisted by accurate automated approaches. The Vanderbilt-EHR algorithm is the first to comprehensively phenotype SCD and other hemoglobinopathies involving variant hemoglobin beta alleles.

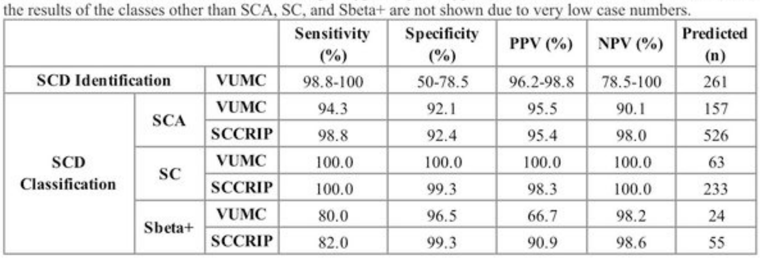
Disclosures Saraf: *Agios:* Consultancy, Other: Advisory board; *BEAM Therapeutics:* Consultancy, Other: Advisory board; *Novartis:* Consultancy, Other: Advisory board, Research Funding; *Forma Therapeutics:* Consultancy, Other: Advisory board, Research Funding; *GBT/Pfizer:* Consultancy, Other: Advisory board, Research Funding, Speakers Bureau. **Weiss:** *Forma Therapeutics:* Consultancy; *Dyne:* Consultancy; *Graphite Bio:* Consultancy; *Novartis:* Consultancy; *Cellarity Inc.:* Consultancy. **DeBaun:** *Novartis, Forma, Vertex:* Consultancy, Other: Consulting.

Table 1. SCD genotypes that are predicted by the hemoglobinopathy algorithm and ids of variants used to validate the predictions



SCD Predicted Genotype	Description	Validation Variant IDs
SS or S Beta Thalassemia 0 (SB0)	A severe form of SCD known as sickle cell anemia (SCA) where a person has two copies of the hemoglobinbeta allele that results in the formation of hemoglobin S or one copy of that allele in addition to a hemoglobinbeta allele that results in little to no production of hemoglobin A	SS –homozygous for rs334 SB0 –heterozygous for rs334 and heterozygous for rs11549407, rs33910569, rs33913712, rs33914668, rs33922842, rs33930702, rs33930977, rs33941849, rs33943001, rs33945777, rs33946267, rs33950507, rs33952266, rs33953406, rs33956879, rs33959855, rs33960103, rs33969400, rs33969853, rs33971440, rs33971634, rs33974936, rs33979901, rs33982568, rs33986703, rs33991059, rs33995148, rs34120553, rs34218908, rs34282684, rs34466953, rs34477959, rs34502690, rs34533941, rs34548294, rs34563000, rs34716011, rs34831847, rs34843844, rs34856846, rs34889882, rs34937014, rs34960334, rs35133315, rs35165357, rs35171933, rs35225141, rs35238478, rs35323748, rs35348864, rs35371965, rs35383398, rs35395625, rs35456885, rs35477349, rs35497102, rs35532010, rs35619054, rs35619688, rs35662066, rs35684407, rs35699606, rs35699671, rs35755331, rs35894115, rs36029927, rs36107977, rs41443947, rs63749815, rs63749819, rs63749957, rs63749960, rs63749977, rs63750022, rs63750040, rs63750099, rs63750128, rs63750223, rs63750320, rs63750407, rs63750475, rs63750513, rs63750532, rs63750556, rs63750655, rs63750692, rs63750774, rs63750783, rs63750842, rs63750860, rs63750915, rs63751076, rs63751152, rs63751201, rs63751218, rs63751306, rs63751478, rs80356820, rs267607291, rs267607293, rs267607295, rs267607297, rs267607298, rs281864574, rs281864898, rs281864899, rs281864900, rs281864901, rs281864902, rs281864903, rs281864904, or rs281864906
S Beta Thalassemia + (SB+)	A form of SCD where a person has one copy of the hemoglobinbeta allele that results in the formation of hemoglobin S in addition to a hemoglobin beta allele that results in reduced production of hemoglobin A	heterozygous for rs334 and heterozygous for rs33913413, rs33915217, rs33916412, rs33925391, rs33931746, rs33931779, rs33941377, rs33944208, rs33951465, rs33972047, rs33978907, rs33980857, rs33981098, rs33985472, rs33994806, rs34039390, rs34135787, rs34196559, rs34305195, rs34451549, rs34483965, rs34500389, rs34527846, rs34598529, rs34690599, rs34704828, rs34793594, rs34809925, rs34883338, rs34999973, rs35004220, rs35099082, rs35256489, rs35328027, rs35352549, rs35424040, rs35485099, rs35578002, rs35724775, rs35799536, rs35949130, rs63750195, rs63750205, rs63750283, rs63750954, rs63751043, rs63751128, rs63751208, rs281864524, or rs281864905
SC	A form of SCD where a person has one copy of the hemoglobinbeta allele that results in the formation of hemoglobin S and one copy of the hemoglobin beta allele that results in the formation of hemoglobin C	heterozygous for rs334 and heterozygous for rs33930165
SD	A form of SCD where a person has one copy of the hemoglobinbeta allele that results in the formation of hemoglobin S and one copy of the hemoglobin beta allele that results in the formation of hemoglobin D	heterozygous for rs334 and heterozygous for rs33946267
SE	A form of SCD where a person has one copy of the hemoglobinbeta allele that results in the formation of hemoglobin S and one copy of the hemoglobin beta allele that results in the formation of hemoglobin E	heterozygous for rs334 and heterozygous for rs33950507
SCA with Persistent Fetal Hemoglobin (SPFH)	A form of SCD where a person SCA (SS or SB0) in addition to an allele that causes elevated and persistent production of fetal hemoglobin	Still trying to add in this genotype validation, will leave out if not enough time and group SPFH with SCA in the predictions

Table 2. Performance of the Vanderbilt-EHR algorithm on two cohorts. The SCD identification portion of the algorithm was applied only to the Vanderbilt University Medical Center (VUMC) cohort, and a sensitivity analysis was performed to address the indeterminate results. For the SCD classification assessment of both the VUMC and SCCRIP cohorts, the cases with indeterminate genotypes or phenotypes (27 VUMC, 9 SCCRIP) were excluded, and the results of the classes other than SCA, SC, and Sbeta+ are not shown due to very low case numbers.



			Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	Predicted (n)
SCD Identification		VUMC	98.8-100	50-78.5	96.2-98.8	78.5-100	261
SCD Classification	SCA	VUMC	94.3	92.1	95.5	90.1	157
		SCCRIP	98.8	92.4	95.4	98.0	526
	SC	VUMC	100.0	100.0	100.0	100.0	63
		SCCRIP	100.0	99.3	98.3	100.0	233
	Sbeta+	VUMC	80.0	96.5	66.7	98.2	24
		SCCRIP	82.0	99.3	90.9	98.6	55

Figure 1

https://doi.org/10.1182/blood-2023-190521