Machine Learning–Driven Screening for Individuals at High Risk of IDH1/2-Mutant Clonal Hematopoiesis

Rushil Srirambhatla[1], Griffen Mustion[2], Duc Tran[2], Nilanjan Chatterjee[3], Kelly Bolton[2]
[1]Department of Applied Mathematics and Statistics, Johns Hopkins University; [2]Division of Oncology, Washington University School of Medicine; [3]Department of Biostatistics, Johns Hopkins University

Introduction

Clonal hematopoiesis (CH), the expansion of somatic mutations in hematopoietic stem and progenitor cells, increases with age and can serve as a preleukemic state for myeloid neoplasms. Individuals with low blood counts and CH (CCUS) and those with certain mutations, including IDH1/2, are at high risk of progression. We have two **ongoing clinical trials** demonstrating high hematologic response rates (>80%) for ivosidenib and enasidenib in IDH12-mutant CCUS. This motivates prevention-oriented identification before co-mutational burden increases. However, **IDH1/2-mutant CH is extremely rare (~0.03% in individuals >40)**, making universal sequencing impractical. Emerging evidence from our group suggests that IDH1/2-mutant CH shows a prodromal phase prior to transformation characterized by abnormalities in blood counts, inflammation, and cardiovascular and autoimmune disease. This supports clinically guided enrichment strategies for carrier identification.

Methods

Cohort. UK Biobank participants (N = 457,698) with whole-exome sequencing linked to demographics, routine laboratory measures, and ICD-10 diagnoses. IDH1/2-mutant CH carriers (n = 115; VAF ≥ 2%) comprised ~0.03% of the cohort.

Feature selection. Age- and sex-adjusted univariate logistic regression was performed across 1,140 candidate variables, with Benjamini–Hochberg correction to retain significant predictors. The final feature set included 171 variables: 31 blood biomarkers, 12 demographic variables, and 128 ICD-10 codes.

Model development. Three classifiers were benchmarked: elastic net logistic regression, random forest, and XGBoost. To address extreme class imbalance, training used repeated class-balanced subsampling of controls (10:1 control:case per chunk) with 3-fold stratified cross-validation. Hyperparameters were tuned by grid search optimizing F1 score.

Ensembling. For each algorithm, the top 100 models were aggregated by averaging predicted probabilities. The decision threshold was tuned via out-of-fold precision–recall F1 optimization on training data.

Evaluation. Performance was assessed on a held-out 30% internal test split by quantifying enrichment of IDH1/2-mutant CH prevalence among the highest predicted-risk percentiles relative to baseline cohort prevalence.

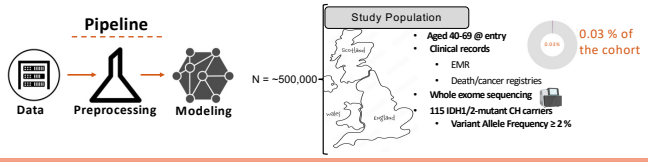
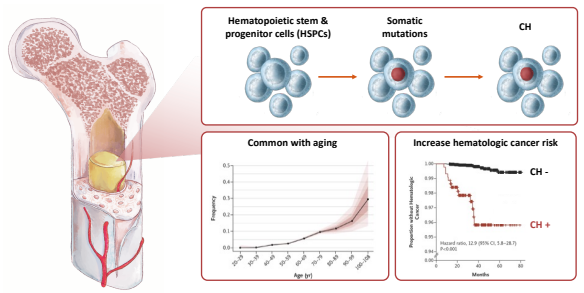


Figure 1. Overview of clonal hematopoiesis: origin, age-dependent prevalence, and association with hematologic cancer risk - Adapted from Jaiswal et al., NEJM 2014



A Elastic-Net: Sample Size & Enrichment by Cut-off

Cut-off (%)	Sample size	Enrichment (fold)
99.9	**113**	**33.1**
99.5	562	20.0
99.0	1,124	13.3
97.5	2,809	6.7
90.0	11,234	4.0
50.0	56,167	1.7
25.0	84,250	1.3
0.0	112,334	1.0

B Random-Forest: Sample Size & Enrichment by Cut-off

Cut-off (%)	Sample size	Enrichment (fold)
99.9	**137**	**56.9**
99.5	682	28.6
99.0	1,364	25.7
97.5	3,410	14.9
90.0	13,639	4.6
50.0	68,193	1.7
25.0	102,289	1.2
0.0	136,386	1.0

C XGBoost: Sample Size & Enrichment by Cut-off

Cut-off (%)	Sample size	Enrichment (fold)
99.9	**137**	**28.4**
99.5	682	17.1
99.0	1,364	14.3
97.5	3,410	13.7
90.0	13,639	4.9
50.0	68,193	1.7
25.0	102,289	1.2
0.0	136,386	1.0

Figure 3. Sample size and fold-enrichment of IDH1/2-mutant CH prevalence across predicted-risk percentile cut-offs on the 30% held-out test set for **(a)** elastic net logistic regression, **(b)** random forest, and **(c)** XGBoost. At each cut-off, sample size reflects the number of individuals retained above the threshold, and enrichment is computed as the prevalence of IDH1/2-mutant CH within that subset relative to the cohort baseline (~0.03%). At the most stringent cut-off (top 0.1%), **random forest achieved the highest enrichment (56.9×)**, followed by elastic net (33.1×) and XGBoost (28.4×), demonstrating that pre-screening can concentrate carriers into a small, sequencing-tractable fraction of the population.

Figure 4. IDH1/2-mutant CH prevalence as a function of predicted-risk percentile cut-off for elastic net, random forest, and XGBoost classifiers on the 30% held-out test set.

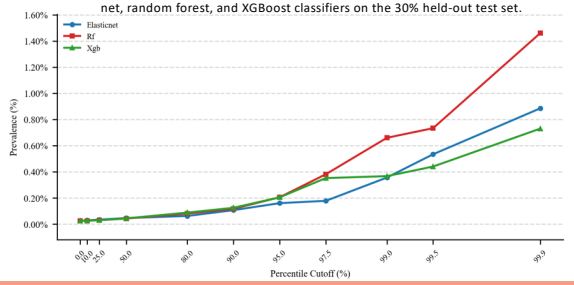
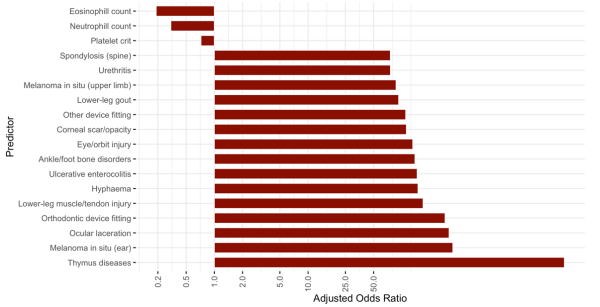


Figure 2. Top predictors of IDH1/2-mutant clonal hematopoiesis identified by age- and sex-adjusted univariate logistic regression, ranked by adjusted odds ratio (log scale).



Results

Predictor discovery. Pre-screening identified significant non-hematologic predictors of IDH1/2 mutation status alongside expected blood biomarkers, most notably thymus diseases and melanoma, as well as signals across trauma/injury, inflammatory conditions, skin neoplasms, and immune/endocrine dysregulation categories.

Model performance. All three classifiers substantially enriched IDH1/2-mutant CH prevalence in their highest predicted-risk strata relative to the ~0.03% cohort baseline. At the top 0.1% risk percentile cutoff, random forest achieved the strongest enrichment (1.47% prevalence), followed by elastic net (0.89%) and XGBoost (0.73%).

Enrichment. Restricting targeted sequencing to the top 0.1% of individuals by model-predicted risk yielded a 57-fold enrichment in IDH1/2-mutant CH yield over baseline for random forest, 33-fold for elastic net, and 28-fold for XGBoost.

Discussion

These results demonstrate that a ML pre-screening layer can concentrate IDH1/2-mutant CH carriers into a small, tractable fraction of the population, making targeted sequencing for early identification of high-risk clones substantially more feasible than universal screening. Converting a ~0.03% population prevalence into a ~1.5% yield within the top risk percentile reframes IDH1/2 CH detection from a prohibitive screening problem into a targeted clinical workflow. External validation in independent biobanks (MGH, Mayo, BioVU, Mexico City Prospective Study), integration of All of Us to improve ancestry representation, and extension to additional CH drivers (TP53, DNMT3A, TET2, SRSF2, ASXL1) are ongoing.

Contact

Rushil Srirambhatla
Johns Hopkins University
Email: rsriram2@jh.edu
Phone: (971) 340-8002

References

1. Jaiswal S, Fontanillas P, Flannick J, et al. Age-related clonal hematopoiesis associated with adverse outcomes. N Engl J Med. 2014;371(26):2488-2498.
2. Genovese G, Kähler AK, Handsaker RE, et al. Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. N Engl J Med. 2014;371(26):2477-2487.
3. Sudlow C, Gallacher J, Allen N, et al. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. 2015;12(3):e1001779.
4. Stein EM, DiNardo CD, Fathi AT, et al. Molecular remission and response patterns in patients with mutant-IDH2 acute myeloid leukemia treated with enasidenib. Blood. 2019;133(7):676-687.
5. DiNardo CD. Mutations in AML: prognostic and therapeutic implications. Hematology Am Soc Hematol Educ Program. 2016;2016(1):348-355.
6. Von Beck K, et al. Clonal hematopoiesis as a pre-malignant state and trajectory to hematologic cancer. Blood Cancer J. 2023.

© Poster Template by Genigraphics® 1.800.790.4001 www.genigraphics.com