A large-scale data simulation platform isolates tumor signal from cell-free DNA and improves tissue of origin prediction accuracy

Kade Pettie, Shiva Farashahi, Jackson Killian, Dorna Kashef, Josh Hubbell, Feras Hantash, Jocelyn Charlton, and Kieran Chacko
Harbinger Health, Cambridge, MA USA

AACR: Liquid Biopsy November 2024 | Poster B065

BACKGROUND

Cell-free DNA (cfDNA) liquid biopsy is a promising non-invasive method for disease detection, but data availability and the complexity and heterogeneity of cfDNA composition pose barriers to model development. Cancer prediction models trained on cfDNA data from clinically-collected blood samples often struggle to isolate tumor-derived signals from confounding factors, and tissue of origin (TOO) models suffer from small sample sizes of rarer tumor types. To address these challenges, we developed a platform to generate diverse, large-scale, high-fidelity simulated cfDNA datasets with matching confounding variable distributions between non-cancer and cancer samples for robust and accurate machine learning (ML) model training.

METHODS

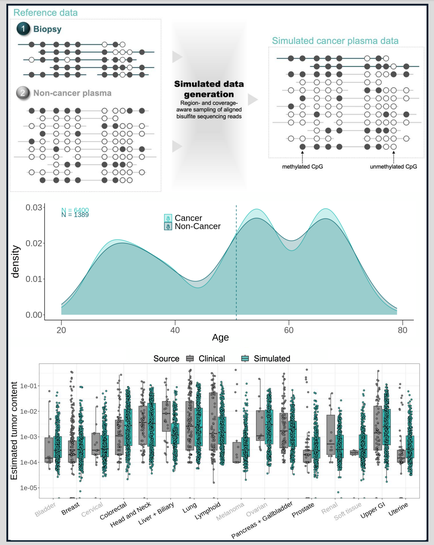

Figure 1: Simulated non-cancer and cancer plasma is age-matched and clinically relevant. (Top) Schematic of simulated cancer generation for one region. (Middle) Age distributions of simulated samples. (Bottom) TC distributions of clinical and simulated cancers (bold = TOO target indication grouping).

Starting with aligned bisulfite sequencing reads from a custom target hybrid capture panel for all samples, we first synthesized non-cancer samples within three age groups by stochastically sampling reads from pairs of non-cancer reference cfDNA samples and combining them in a region- and coverage-aware manner. We then simulated tumor DNA shedding into circulation from 502 reference biopsy samples across 16 indications, optimizing for final tumor content (TC) distributions ranging from 0.001 to 14.2%. This approach expanded a non-cancer dataset of 177 samples to 1,212 simulated samples and generated 6,400 simulated cancer cfDNA samples matching the non-cancer samples in age distribution (**Fig 1**).

RESULTS

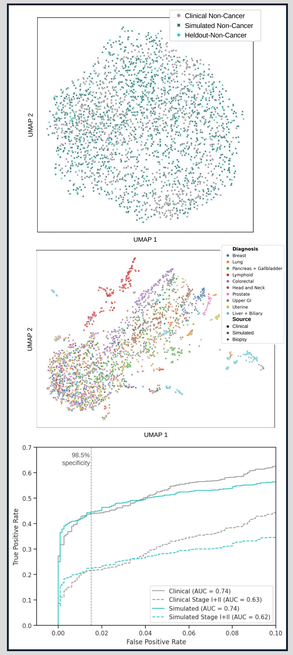

Figure 2. Clinical vs simulated data comparison. UMAPs of non-cancers (Top) and cancers (Middle) using 5 methylation metrics per region. (Bottom) Performance of clinical- and simulated-trained models on clinical samples.

Equivalency

We then assessed this dataset for equivalence with clinical data and ability to improve cancer prediction. Non-linear dimensionality reduction and clustering of methylation signal (quantified by 5 unique metrics per region) showed simulated samples cluster indistinguishably from 2,290 samples (1,149 non-cancer; 1,141 treatment-naïve cancer) from a clinical study (NCT05435066). A binary cancer prediction model trained on the simulated dataset and evaluated on the clinical dataset showed comparable performance to an analogous model trained and cross-validated on the clinical data alone (Benchmark), and outperformed at high specificities, especially for early-stage cancers (**Fig 2**).

Signal tuning

The model trained on simulated data showed reduced reliance on age-associated signal (non-cancer R^2=0.14) relative to the clinical Benchmark model (R^2=0.61). Samples given higher probability scores by the simulated-trained model tended to be younger and have higher estimated TC across both false and true positives (as determined by the Benchmark model), suggesting increased tuning toward tumor content and away from age (**Fig 3**).

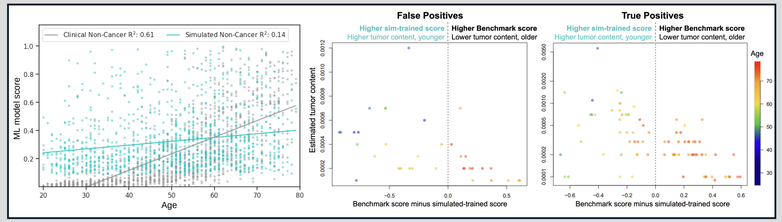

Figure 3. Simulated data reduces model reliance on age. All plotted points are clinical cfDNA samples.

RESULTS

Tissue of origin performance

To evaluate the benefit of these data for TOO prediction, we trained multiclass TOO models using our tumor biopsy reference data with and without the addition of our simulated data for 10 target indication groupings and benchmarked their performance on our clinical dataset. Training with both data types yielded a 10% increase in balanced accuracy. Moreover, peak performance was achieved at different training sample numbers per indication, illustrating our simulated data platform's potential to estimate sample size requirements during clinical study design (**Fig 4**).

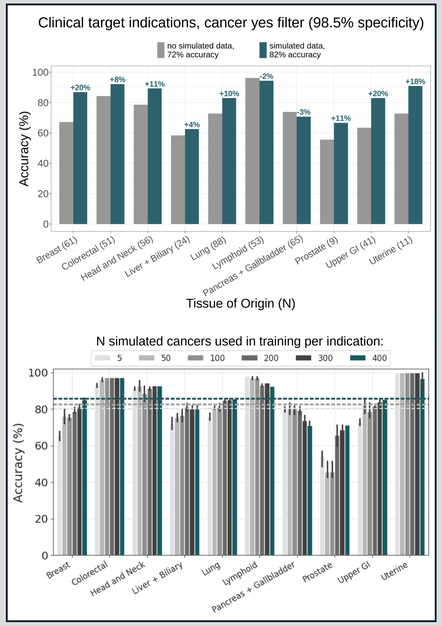

Figure 4: Training with simulated data improves TOO performance. (Top) Barplot of per-indication accuracy among true positives from the clinical data-trained Benchmark binary cancer prediction model when training with clinical data only versus with both data types. (Bottom) Per-indication accuracy when training with increasing numbers of simulated samples (dotted line = mean across indications).

CONCLUSIONS

Our platform generates simulated data that can train ML models to identify tumor-specific biomarkers typically obscured by technical and demographic biases in real-world data. The age-balancing approach is readily applied to other confounders (e.g., sex, ethnicity) to help models learn true disease signatures. The platform is also adaptable to other diseases and biofluids (e.g., Alzheimer's disease, urine), allowing for extension to diagnostic and screening development beyond cancer and blood across the care spectrum.

Disclosure: KP, SF, JK, DK, JH, FH, JC, and KC are full time employees of Harbinger Health, Inc.