5592

SWEDISH CANCER INSTITUTE

Machine learning approach to personalized medicine in breast cancer patients: development of data-driven, personalized, causal modeling through identification and understanding of optimal treatments for predicting better disease outcomes

Kaplan HG[1], Berry AB[1], Rinn KJ[1], Ellis ED[1], Birchfield GR[1], Wahl TA[1], Liu X[1], Tameishi M[1], Beatty JD[1], Dawson PL[1], Mehta VK[1], Holman A[1], Atwood MK[1], Alexander S[1], Bonham C[1], Summers L[1], Khalil I[2], Hayete B[2], Wuest D[2], Zheng W[2], Liu Y[2], Wang X[2], Brown TD[1].
[1]Swedish Cancer Institute, Swedish Medical Center, Seattle. [2]GNS Healthcare, Cambridge.

GNS HEALTHCARE

BACKGROUND

In the era of personalized medicine, a major challenge is harnessing longitudinal data across the cancer care continuum, which includes multi-modal data sets of biological, molecular, and clinical information about patients (pts) and their tumors. There is a growing need for new computing analytics, such as machine learning - an important tool in healthcare bio-informatics. We report our approach to building cancer disease models in an unbiased manner through utilization of a causal machine learning and simulation platform.

METHODS

The Swedish Cancer Institute (SCI) Personalized Medicine Research Program (PMRP) is a prospective registration protocol with the objective of establishing a centralized longitudinal, molecular, and phenotypic data repository. Since 2014, over 1,070 pts have been enrolled, having undergone next generation sequencing (NGS) profiling of their tumors. Of these pts, we identified 100 breast cancer pts who also have detailed longitudinal clinical annotation within our SCI Breast Cancer Registry (BR). The BR was established in 1990 and has over 17,000 cases. BR follows all newly diagnosed breast cancer pts treated at SCI. Detailed information is collected on demo-graphics, clinical presentation, tumor characteristics, staging, treatment and follow-up. The biomarker distribution and time in years from breast cancer diagnosis and NGS test for 100 pts are summarized in **Fig. 1** and **Fig .2**.

All de-identified data, variables, and data points in the multi-modal data types are integrated into normalized data frames. A reverse engineering approach, via the Reverse Engineering and Forward Simulation (REFS) platform, is being utilized, focusing on discovering the complex causal mechanisms that determine which therapies will produce the best outcomes for an individual pt. This method goes beyond traditional approaches that rely on data correlations to match treatments to pts. The breast cancer causal model uncovers many of the possible combinations of causal relationships that drive outcomes and enables "***what if***?" simulations of a variety of interventions, across pts, to determine optimal therapies. Performance metrics and model robustness will be explored using a stratified, n-fold (e.g. 10-fold) cross-validation procedure, which is designed to provide an unbiased estimate of model generalization to new observations. (**Fig. 3**)

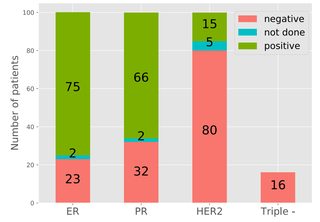


Fig 1. Pts Characteristics - Biomarkers

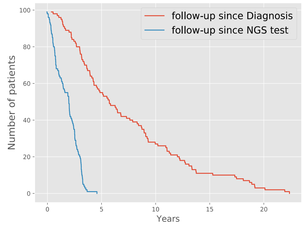
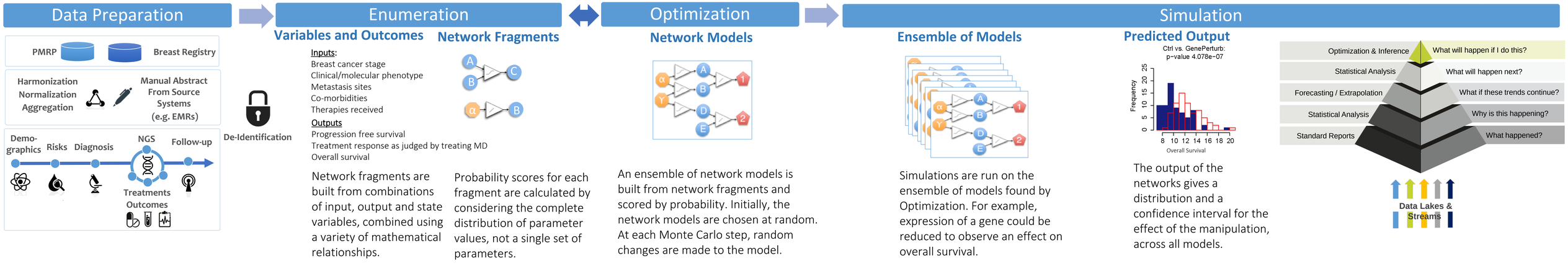


Fig 2. Time from diagnosis and NGS tests

Fig. 3 Breast Cancer Model: Identifying Optimal Treatments in Patients



RESULTS

The causal model and simulations can elevate the providers' abilities to: better understand treatment responses based on pts' unique clinical data and mutational statuses; study different treatment options to optimize management; and understand the complex interactions among variables that lead to a range of treatment outcomes.

Challenges experienced as of today:

- Data access and availability (e.g. system changes over time; lack of access to outside records due to care transfer; lack of EMR automatic download)
- Lack of complete discrete clinical data and reliance on manual curation
- Aggregation of different types of data (e.g. time-series data; static data)
- Prioritizing/balancing granularity vs. breadth of data to guide manual abstraction efforts
- Comparability of data with classification changes (e.g. cancer staging and Her2/FISH)
- Retrospective analysis of cancer biomarkers with phenotype/genotype changes in longitudinal data
- Harmonization and normalization of multiple databases (e.g. coding and data classification)
- Timing of NGS testing vs. clinical application

CONCLUSIONS

Machine learning could potentially provide novel insights into personalized medicine. There are challenges to creating a large enough clean data set that when analyzed will produce results that can be confidently used to inform patient care.

Knowledge generated from the simulations of the disease model can potentially streamline and support the clinical decision-making process, to include molecular tumor board deliberations, and ultimately assist providers in arriving at optimal treatment recommendations for pts.

ACKNOWLEDGEMENTS

We thank CellNetix Pathology and Laboratories for providing majority of tumor genomic profiling.