DOPE: a predictive multi-omic model for cancer prognosis

Cancan Huang, Reetam Ganguli, Laura McCallion, Anessa Petteruti
{cancan_huang, ritambhara_singh,
laura_mccallion, anessa_petteruti}@brown.edu

Brown University — October 12, 2021

Literature Review

Cancer, a notorious threat for global humans, has been ranked as the second leading cause of death with about 1 out of 10 adults in the United States been diagnosed with cancer[1, 2]. Among different cancer types, breast cancer and ovarian cancer are two of the most lethal gynecologic malignancy, despite multi-years of studies in understanding their mechanisms and existing therapies[3, 4]. Furthermore, it has been reported that up to 20% of patients with either breast cancer or ovarian cancer have a relative with one of thse diseases, indicating the potential correlation between these two cancer types[5].

One important goal for current cancer prognosis is to provide guidance for treatment to achieve better survival rate on the basis of patients' clinical profile. Berkson and Gage[6] firstly computed the Life Table to obtain the enhanced frequency distribution of patients' survival times. However, it cannot determine the effects of certain variables in the survival times as it treated patients data with no discrimination. Multiple linear regression models are applied for analyzing survival data but impeded by the problems related to missing censored data and not normal distribution of the data[7]. To tackle the censored data problem, Kaplan and Meier[8] derived a nonparametric method capable of constructing the survival curve. Furthermore, Cox[9] proposed the Proportional Hazards (PH) regression model to reveal the importance of covariates on existing cases. While those statistical methods have been widely used, most of them are mainly focues only on specific clinical data, such as cancer types, cancer diagnosis and etc.

In recent years, researches aiming at early diagnosis and curing these diseases have been fueled by many advances in experimental techniques, which output high throughput and high dimensional multi-omics data of patient samples[3, 10]. With the development of experimental techniques and the emerging abundant data, for example, genomics data (i.e., whole genome data), expression data (i.e., mRNA data) and epigenetic data (i.e., chromosomal modifications), it is necessary and demanding for developing efficient and effective computational methods capable of understanding and handling those multi-omics data for more accurate cancer prognosis[11]. On the other hand, with the burst of computational power and rapid advancement in the technology of artificial intelligence, it is believed that the use of state of the art computational approaches will greatly benefit and acclerate the design and development of novel diagnostic method for breast and ovarian cancers.

To solve these problems and take advantage of the massive data, other methods, including machine learning techniques are applied. Alexe et al.[12] combined principal components analysis (PCA) and k-clustering for breast cancer progression analysis. By applying their method on public microarray breast cancer dataset, it can find clusters and gene markers in the data, which helped identify subtypes of breast cancer. Xu et al.[13] adopted support vector machine (SVM) for breast cancer prognosis on gene expression dataset. They discovered a 50-gene signature with a superior performance than the widely used 70-gene signature in accuracy, sensitivity and speficity. Graf et. al. use genetic association study find the strong correlation between the copy number variation (CNV) signatures and the ovarian cancer from 564 patients[4]. Other machine learing approaches, such as bayesian networks, decision trees and semi-supervised learning, have also been applied in cancer prognosis prediction and shown good performance[14].

Deep learning, a branch of machine learning, shows promise in excellent performance in recent years due to the establishment of public accessible large-scale cancer databases as well as breakthroughs in model architectures[11]. For example, The Cancer Genome Atlas (TCGA) database contains both clinical and molecular data from over 11,000 tumor patiens covering 33 different cancer types[15]. Ching et al.[16]

developed fully-connected neural network model for predicting survival time. The model is trained on TCGA gene expression data, clinical data as well as survival data and achieved better performance than Cox methods and random forest.

References

[1] Kehinde Aruleba, George Obaido, Blessing Ogbuokiri, Adewale Oluwaseun Fadaka, Ashwil Klein, Tayo Alex Adekiya, and Raphael Taiwo Aruleba. Applications of computational methods in biomedical breast cancer imaging diagnostics: A review. *Journal of Imaging*, 6(10):105, 2020.

[2] Rebecca L Siegel, Kimberly D Miller, and Ahmedin Jemal. Cancer statistics, 2019. *CA: a cancer journal for clinicians*, 69(1):7–34, 2019.

[3] Richard Wooster and Barbara L Weber. Breast and ovarian cancer. *New England Journal of Medicine*, 348(23):2339–2347, 2003.

[4] Ryon P Graf, Ramez Eskander, Leo Brueggeman, and Dwayne G Stupack. Association of copy number variation signature and survival in patients with serous ovarian cancer. *JAMA Network Open*, 4(6):e2114162–e2114162, 2021.

[5] M Patricia Madigan, Regina G Ziegler, Jacques Benichou, Celia Byrne, and Robert N Hoover. Proportion of breast cancer cases in the united states explained by well-established risk factors. *JNCI: Journal of the National Cancer Institute*, 87(22):1681–1685, 1995.

[6] GAGE RP et al. Calculation of survival rates for cancer. In *Proceedings of the staff meetings. Mayo Clinic*, volume 25, pages 270–286, 1950.

[7] Farid E Ahmed, Paul W Vos, and Don Holbert. Modeling survival in colon cancer: a methodological review. *Molecular Cancer*, 6(1):1–12, 2007.

[8] Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

[9] David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

[10] Nicolas Goossens, Shigeki Nakagawa, Xiaochen Sun, and Yujin Hoshida. Cancer biomarker discovery and validation. *Translational cancer research*, 4(3):256, 2015.

[11] Wan Zhu, Longxiang Xie, Jianye Han, and Xiangqian Guo. The application of deep learning in cancer prognosis prediction. *Cancers*, 12(3):603, 2020.

[12] G Alexe, GS Dalgin, S Ganesan, C Delisi, and G Bhanot. Analysis of breast cancer progression using principal component analysis and clustering. *Journal of biosciences*, 32(1):1027–1039, 2007.

[13] Xiaoyi Xu, Ya Zhang, Liang Zou, Minghui Wang, and Ao Li. A gene signature for breast cancer prognosis using support vector machine. In *2012 5th International conference on biomedical engineering and informatics*, pages 928–931. IEEE, 2012.

[14] Konstantina Kourou, Themis P Exarchos, Konstantinos P Exarchos, Michalis V Karamouzis, and Dimitrios I Fotiadis. Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, 13:8–17, 2015.

[15] Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. The cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemporary oncology*, 19(1A):A68, 2015.

[16] Travers Ching, Xun Zhu, and Lana X Garmire. Cox-nnet: an artificial neural network method for prognosis prediction of high-throughput omics data. *PLoS computational biology*, 14(4):e1006076, 2018.