Supplement to "Personalised Neoadjuvant Therapy Recommendations for Breast Cancer: A Causal Inference Approach using Multi-omics Data"

1 Multi-omics Data

Table 1 shows a list of example feature classes present in the multi-omics data of breast cancer patients [Sammut et al., 2022].

Table 1: Some Features of Breast Cancer Patients

Feature Class	**Attribute**	**Description**
Clinical	Tumour size	Size of cancer tumour
Clinical	Age	Age of the patient at diagnosis
Clinical	Histological subtype	Histological subtype of the tumour
Clinical	HER2 status	Human Epidermal Growth Factor Receptor status
Clinical	ER status	Estrogen Receptor status
DNA	TMB	Tumour Mutational Burden
DNA	HRD score	Homologous Recombination Deficiency score
RNA	PGR expression	Progesterone Receptor expression
RNA	ESR1 expression	Estrogen Receptor expression
RNA	Taxane score	Score indicating Taxane sensitivity
Digital pathology	Lymphocyte density	Number of Tumour-Infiltrating Lymphocytes (TILs) in the tumour
Treatment	Anti-HER2	Use of anti-HER2 drugs
Treatment	Anthracycline	Use of anthracycline drugs

Multi-omics data, including clinical features, digital pathology, genomic and transcriptomic profiles (DNA and RNA sequencing), were utilised to develop a machine learning model for breast cancer treatment. This model integrates multi-omics data to improve predictive capabilities by providing a comprehensive understanding of tumour molecular and cellular landscapes, surpassing traditional risk stratification methods [Hasin et al., 2017]. The study gathered 74 variables from breast cancer patients, encompassing clinical features, tumour genomic landscapes (DNA, RNA), digital pathology, and treatment details [Sammut et al., 2022].

Key clinical features include patient age, tumour size, histology, hormone receptor (ER and HER2) status, lymph node (LN) status, and ER Allred score. Digital pathology, represented by the number of Tumour-Infiltrating Lymphocytes (TILs) in the tumour (Lymphocyte density), offers insights into immune

micro-environments [Sammut et al., 2022]. DNA-related features such as Tumour Mutational Burden (TMB), Homologous Recombination Deficiency score (HRD score), and key gene mutations (PIK3CA, TP53) reflect tumour heterogeneity and guide personalised therapies. RNA-related features include gene expression profiles of hormone receptors, HER2, immune responses, and tumour stemness, which influence treatment strategies and response prediction.

Treatment-related variables detail chemotherapy regimens, including the number of cycles and the specific drugs used in each cycle (taxanes, anthracyclines, and anti-HER2). The variables "any Anthracycline" and "any anti-HER2" indicate whether the patient received anthracycline or anti-HER2 therapy, respectively. These variables help assess the impact of different chemotherapy regimens on treatment response, providing valuable insights into the effectiveness of neoadjuvant therapy plans for breast cancer patients.

2 Algorithm for CTR

Following the CTR framework, we have devised the Causal-based Therapy Plan Recommendation (CTR) algorithm for personalised therapy plan recommendation, as depicted in Algorithm 1.

3 The application of the CTR model for an alternative setting

In practice, Taxane and Anthracycline are administered as a block-sequential regimen. After selecting the number of drugs to be used, the next step is to determine the therapy sequence for patients receiving both Taxane and Anthracycline. Specifically, we need to decide which drug should be administered first and which second. Our CTR model can provide recommendations for the therapy sequence using a similar framework. We denote three combinations of block-sequential Taxane and Anthracycline as follows:

- Therapy Sequence 1 (TS1): Taxane is administered first (T = 1) and Anthracycline second (T = 0), where A = 1 indicates that Anthracycline is administered in any cycle.
- Therapy Sequence 2 (TS2): Anthracycline is administered first (T = 1) and Taxane second (T = 0).
- Therapy Sequence 3 (TS3): Only Taxane is used, with no Anthracycline administered (A = 0).

In this case, TS3 is identical to Therapy Plan 3 in the previous experiment setting which use only Taxane, allowing us to use it for cross-verification. For a breast cancer patient receiving recommendations from both models, TP3 and TS3, it provides additional confidence in applying only Taxane as the therapy plan.

Algorithm 1 Causal-based Therapy Plan Recommendation (CTR)

— **Training Causal Tree Models** —

Input: Training dataset $\mathbf{D}$ for a set of attributes $\mathbf{X}$, therapy plans $\mathbf{T}$ and outcome Y.

Output: Causal Trees set $\mathbf{Z} = \{Z_1, Z_2, \ldots, Z_q\}$.

0: Let $\mathbf{Z} =$
0: **for** each Therapy Plan $TP \in \mathbf{T}$ **do**
0: Fit a Generalised Linear Model (GLM) to estimate propensity score.
0: Call the Causal Tree method with TP as the treatment variable, outcome Y and $\mathbf{X}$ as the covariate set.
0: The obtained causal tree is stored in $\mathbf{Z}$ as Z_k.
0: **end for**
return Causal trees set $\mathbf{Z}$ for each Therapy Plan. =0

— **Personalised Recommendation** —

Input: Dataset $\mathbf{D}_R$ for a set of attributes $\mathbf{A}$, Causal trees set $\mathbf{Z} = \{Z_k\}$.

Output: Recommendation set $\mathbf{R}(\mathbf{A}, TP, CATE_{TP})$, where TP is therapy plan, and $CATE_{TP}$ is the estimated conditional average treatment effect.

0: Let $\mathbf{R} =$
0: **for** each individual $i \in \mathbf{D}_R$ **do**
0: Retrieve attribute values $\mathbf{A}(i)$.
0: **for** each causal tree $Z_j \in \mathbf{Z}$ **do**
0: Retrieve causal factor TP_k of causal tree Z_k.
0: **if** value of attribute $TP_k(i) = 0$ **then**
0: Search for the subset that individual i belongs to.
0: Retrieve $CATE_{i(TP_k)}$.
0: **end if**
0: **end for**
0: Select the therapy plan TP with the largest $CATE_{i(TP)}$ to recommend for individual i.
0: Add tuple $(\mathbf{A}(i), TP, CATE_{i(TP)})$ into $\mathbf{R}$.
0: **end for**
return $\mathbf{R}$. =0

We also compared this experimental setting to the current protocol. The results in Figure 1 show that our model outperforms the current protocol in terms of Recovery Rate, and Figure 2 demonstrates superior AUUC performance compared to all baseline methods.

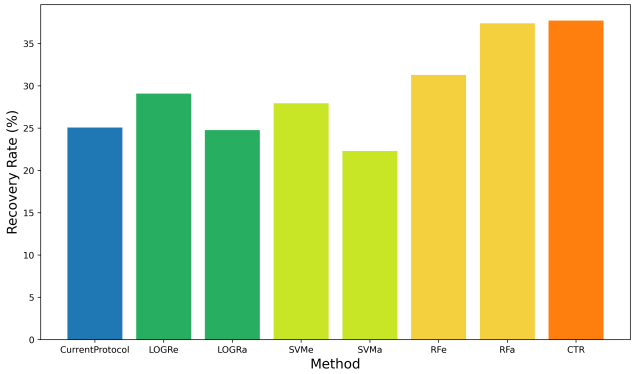

Figure 1: Comparison of Recovery Rates for Therapy Sequence setting

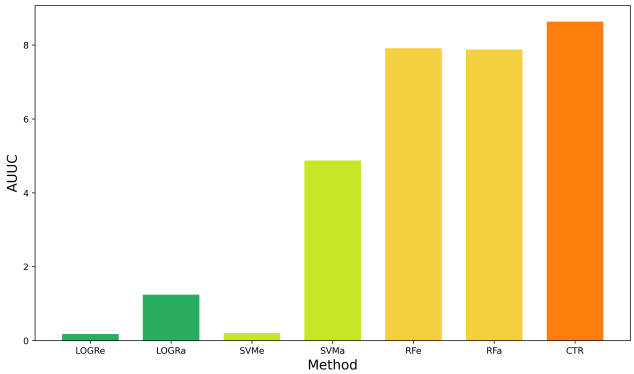

Figure 2: Comparison of Model Performance for Therapy Sequence Setting

References

Y. Hasin, M. Seldin, and A. Lusis. Multi-omics approaches to disease. *Genome biology*, 18:1–15, 2017.

S.-J. Sammut, M. Crispin-Ortuzar, S.-F. Chin, E. Provenzano, H. A. Bardwell, W. Ma, W. Cope, A. Dariush, S.-J. Dawson, J. E. Abraham, et al. Multi-omic machine learning predictor of breast cancer therapy response. *Nature*, 601(7894):623–629, 2022.