Genetic profiling to predict recurrence of cervical cancer

Final Presentation - PH240B

Cameron Adams

December 30, 2020

Outline

- Background
- Summary of Analysis Plan
- Road Map
  - The Data
  - The Model
  - The Target Parameter
  - Identification, eIC, TMLE
- Stage 1+2 methods
- Results
- Limitations and Next steps
- Conclusions

Background

Genetic profiling to predict recurrence of early cervical cancer

Yoo-Young Lee [1], Tae-Joong Kim, Ji-Young Kim, Chel Hun Choi, In-Gu Do, Sang Yong Song, Insuk Sohn, Sin-Ho Jung, Duk-Soo Bae, Jeong-Won Lee, Byoung-Gie Kim

Affiliations + expand

PMID: 24145113 DOI: 10.1016/j.ygyno.2013.10.003

NCBI GEO: GSE44001

- Existing prognostic models that predict tumor recurrence following treatment using clinical characteristics are not widely used
- There is debate about relevant characteristics and variability in their measures
- Gene expression data can be used to a confirm/identify cancer staging and to provide prediction of clinical outcomes in patients
- Authors used CoxPH lasso to identify gene-set model and predicted survival from tumor recurrence at 120 months (KM + log-rank)

My Goal: Use causal methods to identify gene set model and estimate treatment specific survival among high and low risk groups.

Background

Available Data

	No Recurrence	Recurrence
n	262	38
T, months (med [IQR])	148.5 [80.3, 218.8]	88 [28.3, 161]
T>120 months	156 (59.5)	16 (42.1)
Cancer stage (%)		
IA2	13 (5.0)	0 (0.0)
IB1	196 (74.8)	21 (55.3)
IB2	19 (7.3)	9 (23.7)
IIA	34 (13.0)	8 (21.1)
max. diam (cm)	25.12 (13.57)	34.55 (9.72)

$p = 29,377$ normalized gene expression measurements
Note: Survival times have different distribution than reported in manuscript

Summary of Analysis Plan

Two stages to project:

1. Stage 1: Identification of high/low risk groups using gene expression data
   - sl3 for estimate of survival $>$ 120m months given gene expression values
     - IPCW via KM
     - Screening (X: 29K $\rightarrow$ 299)
     - sl3 for survival for $t > 120$ months
     - Obtain predictions for each individual, $\hat{Y}$
   - Define High/Low risk "treatment" groups
   - High-risk: $\hat{Y} <$ median($\hat{Y}$)
   - Variable Importance
2. Stage 2: Estimation of treatment specific survival/hazard among high and low risk groups up to 120 months
   - sl3 via "long-data" method
   - survtmle

The Data

$O = (\tilde{T} = min(T, C), \Delta, X, W)$

- $O \sim P_0$
- n *i.i.d* draws from $O : o_1, ..., o_n$

Current status failure time with right censoring

- T : Tumor recurrence at time t during follow-up
- C : Monitoring time
- $\Delta = \mathbb{I}(T \leq C)$
- $\tilde{T} = min(T, C)$
- W : Baseline clinical covariates (Stage, diameter)
- X : Normalized gene expression measurements from tumor sample (p=29K)

The Data

Convert into longitudinal structure:

$$dN(t), dA(t) : (W, X(dN(t), dA(t) : t))$$

- $dN(t) = I(\tilde{T} \leq t, \Delta = 1)$ # of observed failures at time t
- $dA(t) = I(\tilde{T} \leq t, \Delta = 0)$ # of observed censoring events at time t
- W: matrix of baseline covariates
- X: matrix of gene expression measurements
- A: "High" vs. "Low" risk treatment groups from stage 1 prediction model

The Model

$\mathcal{M}$

- $P_0 \in \mathcal{M}$
- Non-parametric
- CAR: $\mathbb{P}(C \geq t \mid X, W)$ assuming $C \perp T$

The Target Parameter

$\Psi : \mathcal{M} \rightarrow \mathbb{R}$

Stage 1: $\Psi^{X}(P) = P(T > t_0, \Delta = 1|X)$ at $t_0 = 120$ months

Stage 2: Treatment specific survival of tumor recurrence at $t_0 = 1, 2, ..., 120$ months

$\Psi^{a,W}(P) = P(T > t_0, \Delta = 1|A = a, W)$

Loss Function: Binomial log-likelihood

$$L_{loglik}(O, \Psi) = \sum_{t} I(\tilde{T} \geq t) log(\psi(t|W))^{dN(t)} log(1 - \psi(t|W))^{dN(t)}$$

Identification

Sequential randomization assumption:

- Backdoor paths are blocked by adjustment for covariates at each time point

Positivity assumption:

- $P(A = a|W) > 0$

eIC

Factorize likelihood:
$P_0(O = o) =$
$P_0(W)\prod_{t=1}^{t_0} Q_{dN(t),0}(dN(t)|Pa(N(t)))\prod_{t=1}^{t_0} g_{dA(t),0}(dA(t)|Pa(A(t)))$

Initial gradient (IPTW):
$H(G)(o) = \frac{I(\bar{a}=1)}{G(A|X_O)}$
$D(P) = H(G)(o)N_j(t_0) - \Psi(P)$

eIC:
D*(Q,G)(o) = E(D|W) + E[(D|N(t), Pa(N(t))] - E(D|Pa(N(t))

Hat tip: Lauren Eyler Dang, Yunzhe Zhou, Pablo Freyria Duenas

TMLE overview

1. Estimate censoring mechanism and treatment mechanism to obtain G_n and g_n using SL
2. Estimate initial fit of $\bar{Q}_0$ using SL to obtain $\bar{Q}_{k,n}$ with $k = 0$
3. Use logistic regression to estimate ϵ_n:

$$logit(dN_1(t)) = logit(dN_0(t)) + \epsilon_n H_{n0}$$

4. Update $\bar{Q}^*_{k+1,n} = \bar{Q}_{\epsilon_n,k+1}$
5. Iterate process until ϵ_n is sufficiently minimized
6. Obtain TMLE

$$\Psi^*_n = n^{-1} \sum_{i=1}^{n} \bar{Q}^*_n(O_i)$$

Stage 1: Variable Importance

- Outcome: Survival > 120 months
- Data: X: 29K gene expression measurements

1. Create 10-fold CV splits for use in rest of algorithm
2. Estimate IPCW via KM
3. Screening via Lrnr_screener_randomForest
   - nVar=299, ntree=1001, mtry=171
4. Estimate survival, $t_0 > 120$ months, with sl3
   - glm, ridge, elasticnet, lasso, 30 xgboosts
5. CV sl3 for performance of learners
6. varimp with loss_loglik_binomial
7. Get Ψ_n^X
   - "high-risk" - $\Psi_n^X \leq \text{median}(\Psi_n^X)$
   - "low-risk" - $\Psi_n^X > \text{median}(\Psi_n^X)$

Stage 2: Treatment specific hazard

SL estimate

1. Create long-data format, 1 observation for every time point person observed until failure or censoring
2. Learners: glm, ridge, lasso, xgboots
3. Loss: binomial log-likelihood
4. Train model on long-data and get predictions of survival for each time point for each person
5. Take product across all time points with each treatment group to get treatment specific hazard/survival predictions

survtmle

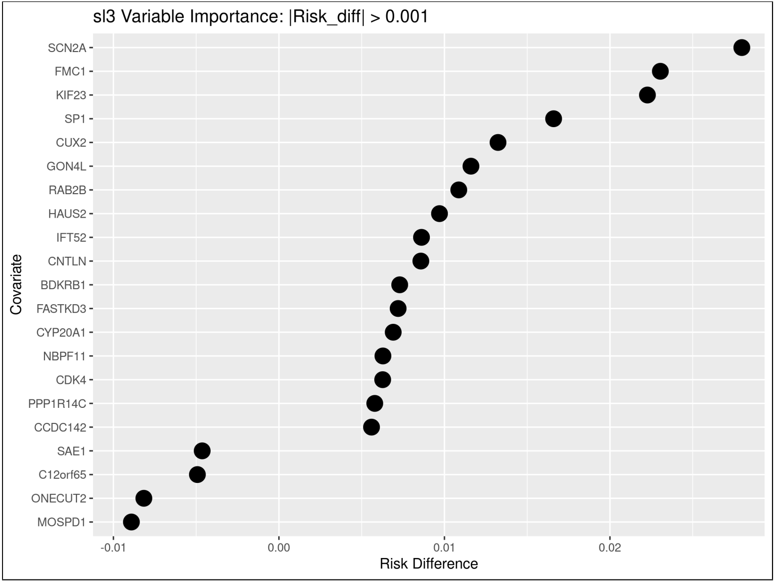
1. Use SL to estimate probability of censoring, treatment, and failure time (glm, xgboost, ridge, lasso)
2. Cause-specific hazards method
3. 10-Fold CV
4. Run at time points of interest, $t_0 = 10, 20, 30, ..., 120$ months
5. Obtain eIC at each time point
6. Estimate simultaneous CI for treatment specific survival/hazard at each time point

Results - Stage 1 SL

- 44 learners
- Honest 10-fold CV
- Coefficient $>$ 0 displayed below

learner	coef.	invlogit(mean_risk)	SE_risk
glm	0.051	0.939	0.538
glmnet_ridge	0.227	0.680	0.512
xgboost_20_1_4_0.1	0.130	0.657	0.510
xgboost_20_1_6_0.1	0.130	0.657	0.510
xgboost_20_1_8_0.1	0.130	0.657	0.510
xgboost_50_1_4_0.01	0.071	0.657	0.511
xgboost_50_1_6_0.01	0.071	0.657	0.511
xgboost_50_1_8_0.01	0.071	0.657	0.511
SuperLearner		0.684	0.510

Results - Stage 1 VIMP



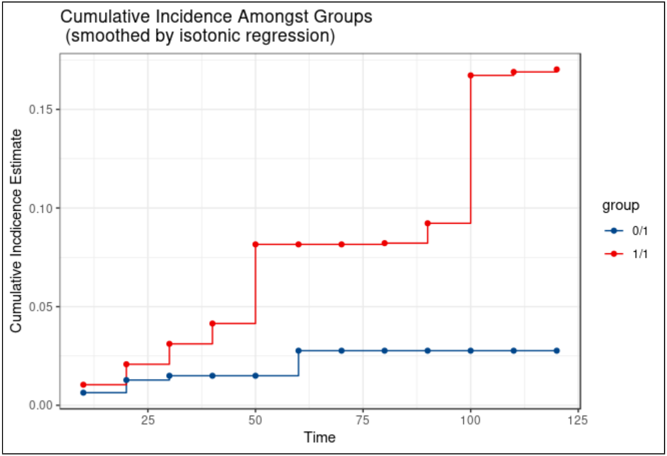
Results - Stage 2 Treatment specific survival

isoreg() used to smooth cumulative hazard point estimates

		Low-risk		High-risk	
t	Method	Haz	95% CI	Haz	95% CI
10	survtmle	0.006	0.0044, 0.0082	0.010	0, 0.032
20		0.013	0.0089, 0.016	0.020	0, 0.067
30		0.015	0.0092, 0.02	0.030	0, 0.084
40		0.015	0.0078, 0.022	0.041	0, 0.1
50		0.015	0.015, 0.015	0.081	0.016, 0.15
60		0.027	0.018, 0.037	0.081	0.016, 0.15
70		0.027	0.016, 0.039	0.081	0.016, 0.15
80		0.027	0.015, 0.04	0.081	0.013, 0.15
90		0.027	0.013, 0.042	0.091	0.017, 0.16
100		0.027	0.027, 0.027	0.169	0.091, 0.25
110		0.027	0.027, 0.027	0.172	0.094, 0.25
120		0.027	0.027, 0.027	0.174	0.096, 0.25
120	sl3-long	0.0028	0.0028, 0.0028	0.131	0.11, 0.15

Results - Stage 2 Treatment specific survival

From **timepoints()**



Results - Stage 2 Treatment specific survival

Comparison between survtmle mansucript survival

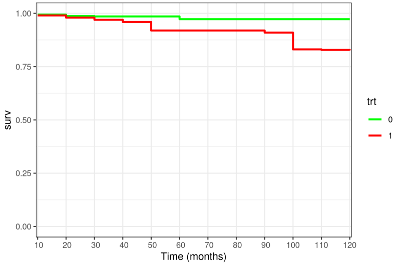
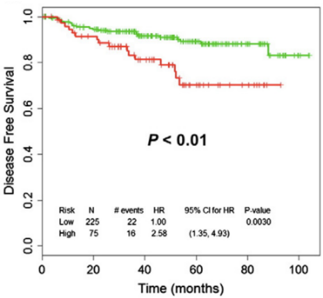


Figure: Survival

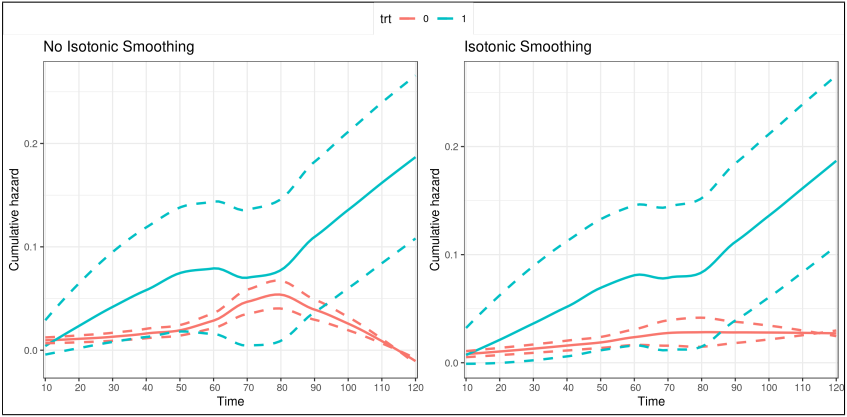


Risk	N	# events	HR	95% CI for HR	P-value
Low	225	22	1.00		0.0030
High	75	16	2.58	(1.35, 4.93)	

Figure: Manuscript

Results - Stage 2 Treatment specific survival

simultaneous CI, points estimates smoothed with **isoreg()**



Limitations and Next Steps

Limitations

- Stage 1 not estimated across all time points, only $t_0 = 120$
- Stage 1 and Stage 2 used same data. Ideally would use different data (training and validation)
- Much of the baseline covariate data used in original article is not publicly available.
    - Residual confounding from lack of baseline data
    - Unable to compare to prediction model to clinical model

Next steps:

- Stage 1 and Stage 2 super learner hazard estimators across more time-points
- Run survtmle across a finer grid of time points (i.e, $1, 2, 3, \ldots, 120$, rather than $10, 20, 30, \ldots, 120$)
- Annotation of top genes from variable importance
- Use genes identified in article to create prediction model and compare to genes discovered here

Overall conclusion

- Gene expression values from cervical tumor samples are predictive of tumor recurrence following treatment (surgery + chemo/radiation) after adjusting for cancer stage and tumor size
- Genes found in Stage 1 are different from genes found by study authors, but that could be due to data issues, rather than different biology. More investigation needed to tease this out.