Using Artificial Intelligence to Improve Capture of Metastatic Breast Cancer Status in Electronic Health Records

Authors: Agrawal S*, Colano V, Chandrashekaraiah P, Vaidya VP, Charest F, Manirevu SVK, Inbar O, Jun MP, Stepanski EJ, Walker MS, Peevyhouse A, Narayanan B, Hyde B

***Contact**: sagrawal@concertohealthai.com
Institution: Concerto HealthAI

Objectives: Though an important prognostic feature in cancer, stage information is often missing from patient's Electronic Health Records (EHRs) and unavailable in claims data. Efforts to identify metastasis status from linked clinical and claims data have previously been made with a limited success [1,2]. The primary objective of this study is to develop and validate an artificial intelligence model that classifies metastatic status in BC patients at their last observed timepoint (proxy for present-day) using previously collected, de-identified, retrospective large scale EHR data.

Methods:

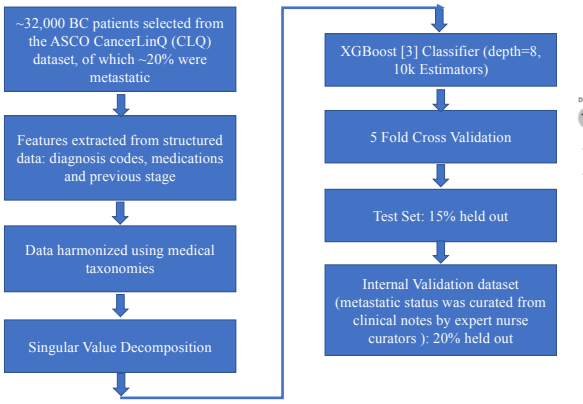

Fig 1: Flowchart explaining the methodology used for modeling, testing and validation

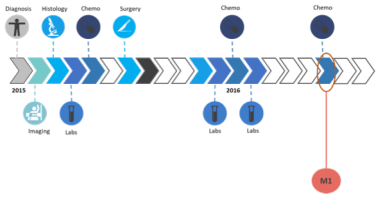

Fig 2: Depiction of a patients journey as captured in the EHR and the features used to build the model

Results: On the full curated validation set, the model had an AUC-ROC of 0.9 and a globally weighted f1 score of 0.87. The model was able to predict metastasis in BC with a positive predictive value (PPV) of 0.93 and negative predictive value (NPV) of 0.86. For the subset of patients with no recorded stage or secondary malignancy information in EHR, the model predicted metastasis with a PPV of 0.89. Compared to business rules alone, the model can correctly identify 32% more metastatic cases.

	True Label Metastatic **Total 2142**	True Label Not Metastatic **Total 4162**
Predicted Metastatic	1494 (TP)	119 (FP)
Predicted Not Metastatic	648 (FN)	4043 (TN)

	Metastatic Class	Not Metastatic Class
Precision	92.6%	86.2%
Recall	69.7%	97.1%

Overall Accuracy =87.8%, AUC = 0.9

Fig 3: Results of validation on expert curated hold out set

		1494 TP Cases	648 FN Cases	119 FP Cases	4043 TN cases
1	Cases with indication of metastasis in EHR	620	201	-	-
2	Cases with indication of NO metastasis in EHR	52	114	51	3639
3	Cases with missing stage and M-stage information	562	145	68	404
4	Cases with ambiguous data (stage 0-3 or M0) and secondary codes present	260	188	-	-

Table 1: Breakup of cases by their complexity based on information present in the EHR

Conclusions: This model yielded high precision and recall, and thus could be an important tool for imputing missing metastatic status information in EHRs. The performance of the model compares very favourably against a similar model created to impute metastatic status in prostate cancer patients [4]. Using this could save substantial time and resources by quickly identifying most eligible candidate patients for clinical trial enrolment or retrospective outcomes studies as compared to expert manual curation.

Future Direction: The model can be further improved by adding more features such as biomarker status, performance status, radiation therapy information etc. as well as by adding rules to catch the small number of easier cases which are currently being missed. We have developed a v2 of this model by incorporating all of these changes which significantly increases the performance. The new model now has an AUC 0f 0.98 with a precision and recall for the metastatic class at 0.98 and 0.88 respectively.

References:

1. Nordstrom BL, Whyte JL, Stolar M, Mercaldi C, Kallich JD. Identification of metastatic cancer in claims data. *Pharmacoepidemiology and Drug Safety*. 2012;21(Suppl 2):21–28
2. Whyte JL, Engel-Nitz NM, Teitelbaum A, Gomez Rey G, Kallich JD. An Evaluation of algorithms for identifying metastatic breast, lung, or colorectal cancer in administrative claims data. *Med Care* 2015;53:e49–57
3. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (ACM, 2016)
4. Seneviratne MG, Banda JM, Brooks JD, Shah NH, Hernandez-Boussard TM. Identifying cases of metastatic prostate cancer using machine learning on electronic health records. *AMIA Annu Symp Proc*. 2018 Dec 5; 2018:1498-1504