Validating a Machine-Learning Approach to Cancer Stage Identification Using Medicare Claims and SEER Data

*MILLIMAN, INC., New York, NY, USA, #EXACT SCIENCES CORPORATION, Madison, WI, USA

OBJECTIVE

- Administrative claims data can provide information about real-world costs, treatments, and mortality for millions of patients with cancer, but claims’ diagnosis codes lack stage information, limiting research capabilities.
- Accurate assignment of cancer stage at diagnosis through claims would expand population research capabilities.
- This work aimed to build and validate a predictive machine-learning algorithm to assign patients’ cancer stage at diagnosis using claims data.

METHODS

Patient Identification

- Patients with incident non-small cell lung cancer (NSCLC), colon cancer (CC), or rectal cancer (RC) diagnosed between 2016-2017 were identified using the SEER-Medicare data.
- In the SEER dataset, NSCLC, CC, and RC tumors diagnosed in 2016 and 2017 were identified using primary tumor codes, histology codes, and the diagnosis year listed in the registry. Patients with another tumor of a different type apparent in 2016-2017 were excluded from the analysis.
- Patients with <1 month of Medicare Parts A/B/D enrollment in 2016-2017, <12 months of A/B/D enrollment prior to the SEER diagnosis date, and cancer-related treatment within one year before index or prior cancer diagnoses were excluded. See Table 1 for Patient Waterfall.

Data Setup and Machine Learning Model Development

- Patients’ claims were flagged for evidence, frequency, and timing of cancer-related surgeries, anti-cancer therapies, and nursing home residence. Claims’ diagnosis codes lack stage information, limiting research capabilities.
- These clinical flags plus age, gender, frailty-related diagnoses, and nursing home residence were tested as predictors of patients’ SEER-derived AJCC stage for each cancer type.
- The flagged patients were analyzed using predictive multinomial logistic regression (nnet package; Venables and Ripley 2002) with R Statistical Software (v4.1.2; R Core Team 2021).
- Patients “actual stage” was based on the derived AJCC stage associated with their primary incident tumor record in SEER.
- The model trained separately on 70% of each cancer sample and tested on 30%.

RESULTS

- The NSCLC (n = 13,494) overall staging accuracy was 77.5% (CI 76.2%-78.8%) (Figures 1A & 1B).
- The Colon Cancer (n = 7,145) overall staging accuracy was 82.3% (CI 80.6%-83.9%) (Figures 2A & 2B).
- The Rectal Cancer (n = 1,424) overall staging accuracy was 67.4% (CI 62.7%-71.8%) (Figures 3A & 3B).
- All models most accurately identified patients with stage 4 disease.

- Among all cancer types analyzed, metastatic diagnoses and surgery occurred sooner after diagnosis for stage 4 cohorts, but chemotherapy was received later after diagnosis when compared to stage 0/1/2 cohorts. Stereotactic ablative radiotherapy (SABR) was more common for stage 0/1/2 cohorts. Table 2 shows key stage predictors for each cancer.

CONCLUSION

- Multinomial logistic regression using claims data accurately predicts stage at diagnosis for patients with NSCLC, CC, and RC. We were able to identify several significant predictors from claims data that can distinguish stage within each cancer type.
- These findings suggest machine-learning algorithms may be a viable approach for assigning patients’ cancer stages at diagnosis when analyzing administrative claims data.
- This work could facilitate widespread analyses of cancer costs by stage and the impacts of early detection and treatment.
- Future validation on more recent data could be useful for incorporating new or emerging treatment advancements into the model.

Table 1. Patient Criteria

<table>
<thead>
<tr>
<th>Criteria</th>
<th>NSCLC</th>
<th>Colon</th>
<th>Rectal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patients in registry with tumor matching cancer type</td>
<td>138,444</td>
<td>79,530</td>
<td>19,160</td>
</tr>
<tr>
<td>2016 or 2017 year of first diagnosis</td>
<td>67,661</td>
<td>37,693</td>
<td>8,986</td>
</tr>
<tr>
<td>Meet Medicare enrollment requirements with no evidence of prior cancer treatment</td>
<td>24,670</td>
<td>13,024</td>
<td>2,628</td>
</tr>
<tr>
<td>SEER-derived AJCC stage available for tumor</td>
<td>13,494</td>
<td>7,145</td>
<td>1,424</td>
</tr>
</tbody>
</table>

Figure 1A. NSCLC Stage Distribution

Figure 1B. NSCLC Staging Accuracy

Figure 2A. CC Stage Distribution

Figure 2B. CC Staging Accuracy

Figure 3A. RC Stage Distribution

Figure 3B. RC Staging Accuracy

Table 2. Key Predictors of Cancer Stage

<table>
<thead>
<tr>
<th>Stage</th>
<th>NSCLC</th>
<th>Colon</th>
<th>Rectal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Days to 1st Chemotherapy Treatment</td>
<td>64%</td>
<td>77%</td>
<td>82%</td>
</tr>
<tr>
<td>Metastatic Diagnoses</td>
<td>86%</td>
<td>72%</td>
<td>89%</td>
</tr>
<tr>
<td>Evidence of Hospice Care</td>
<td>68%</td>
<td>61%</td>
<td>79%</td>
</tr>
</tbody>
</table>

Table 2 shows key stage predictors for each cancer.

Presented at ISPOR 2023 | May 8, 2023 – Poster Session #2