Accurate Oncology Regimen Annotation and analysis of real-world oncology treatment patterns across 5 academic institutions.

Travis Zack, Asieh Golozar, Christian Reich, Atul Butte, Jeremy Warner, Julian Hong

Background

Recently, there have been increasing efforts to use machine learning on large datasets obtained from electronic medical records (EMR) to inform and improve clinical care, yet the standardization and organization of this data has so far limited its utility in oncology. The options and complexity of cancer continue to expand, and with frequent protocol modifications due to patient intolerance, accurate and properly controlled comparisons and cohort identification across providers and institutions can be challenging. This study aimed to create a tool for standardized identification of 1) anti-neoplastic regimen and 2) regimen modification by using an open-source database of anti-neoplastic treatments and applying this database to structured EMR oncology treatment data across 5 academic centers.

Methods

Here we expand on previous methods[1] to leverage HemOnc.org, an open-source, comprehensive database of oncology treatment protocols, to create a database of 5,146 regimens across 146 hematology and oncology diseases that includes information about drug names, optimal dosages, administration days, cycle lengths, and number of cycles within a complete treatment. We use rule-based natural language processing to convert this text database into a structured database of anti-neoplastic regimens. We have developed a convolutional time series maximum likelihood estimate algorithm to identify the most likely regimen a patient is undergoing at each point in a patient's treatment history. This algorithm allows identification of 1) lines of therapy 2) duration of treatment 3) number of therapy cycles. This algorithm also detects deviations from standard therapy and can be used for further analysis of patterns of treatment modifications across and within institutions through unsupervised and semi-supervised techniques.

Preliminary Results

As an initial application for these methods, we have collected drug administration data from 20,000 patients from the UCSF OMOP database (11,000 patients with gastrointestinal (GI) malignancies, 9,000 patients with genitourinary (GU) malignancies). For each of these patients, we have applied the algorithm described above to identify over 200 unique regimens given to these patients (example data Figure 1). This includes identification of complex regimens involving multiple antineoplastic therapies, with high levels of accuracy based on manual validation of results. We plan to apply these methods to the over 100,000 cancer patients treated across the University of California health network, which includes 5 academic institutions and other OHDSI data partners. As a proof of concept, we have shown that using these techniques can uncover differences in practice patterns in the peri-operative administration of systemic chemotherapy for the treatment of localized pancreatic adenocarcinoma (Figure 2)

Conclusion

Cancer treatment involves highly specific and complex regimens that often include combinations of anti-neoplastic pharmaceuticals, which can make analysis of observational data challenging. We introduce a standardized model to extract and label regimen information from complex, real-world patient drug administration data across 145 malignancies. This allows conversion of individual anti-neoplastic drug data into a episodes of oncologic treatment. We plan to apply this model to real-world data (RWD) from a large network of data partners form different settings and geographics with a wide range of malignancies. This approach is an effective strategy for more accurate identification of oncology treatment regimen and thus enabling reliable and large-scale oncology studies. Our preliminary results show the utility of this tool both within and across disease groups and institutions and our upcoming work with highlight the strength of these analyses across all oncology patients treated within the University of California system over the last 10 years.

References

1. Belenkaya, R. *et al.* Extending the OMOP Common Data Model and Standardized Vocabularies to Support Observational Cancer Research. *JCO Clin. Cancer Informatics* 12–20 (2021) doi:10.1200/cci.20.00079.

Figure 1: Schematic process of extracting treatment regimens from individual drugs.

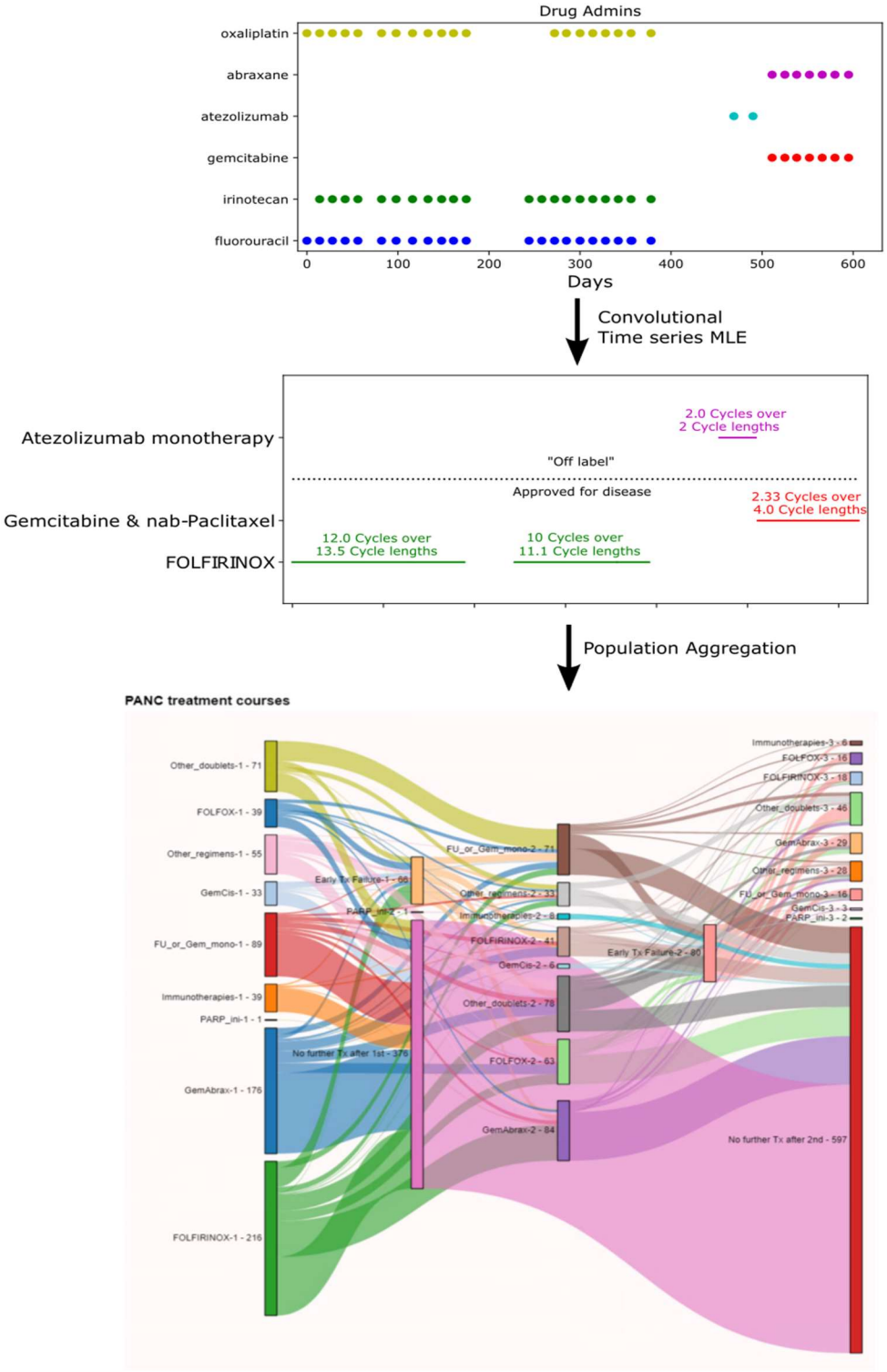
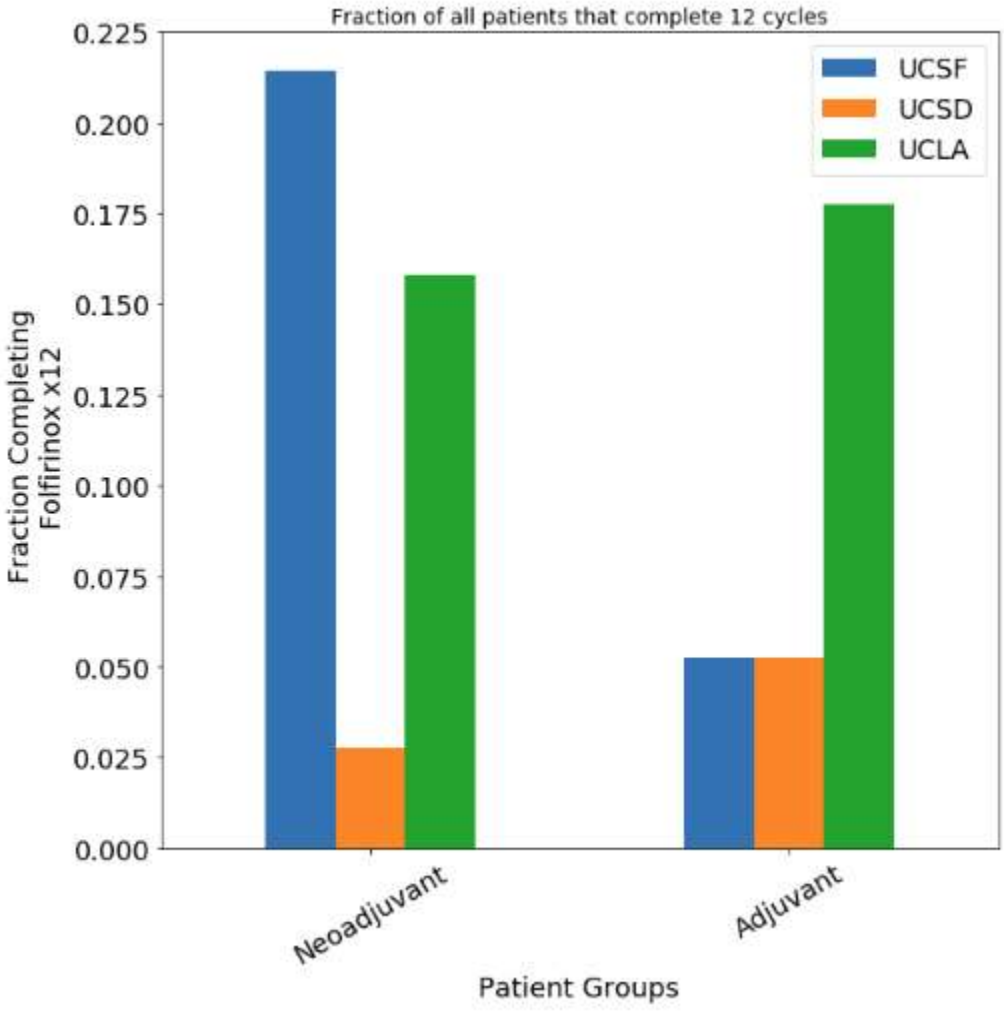


Figure 2: Distribution of patients who completed 12 cycles of Folfirinox in the adjuvant and neo-adjuvant settings for the treatment of localized pancreatic adenocarcinoma



References/Citations