Identifying Eligibility for Chimeric Antigen Receptor T-cell Therapy Among Diffuse Large B-Cell Lymphoma Patients Using Real-World Data and Unsupervised Machine Learning

Yinan Wang MPP[1], Myles S Nickolich MD[2], Charleen Hsuan JD PhD[1], Christopher S. Hollenbeak PhD[1], David J Vanness PhD[1]

1 Department of Health Policy and Administration, Pennsylvania State University; 2 Division of Hematology and Oncology, Penn State College of Medicine

MSR43

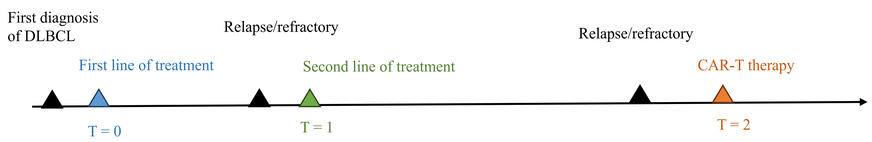
BACKGROUND

- Research on effectiveness and equity implications of advanced cancer treatment using real world data (RWD) requires accurate characterization of treatment patterns.
- Identifying diffuse large B cell lymphoma (DLBCL) patients who are eligible for CAR-T therapy requires correct determination of whether and when switches to new lines of systemic therapy occur due to relapsed or refractory disease.[1] (see Exhibit 1)
- However, many hematological oncology treatments delivered in practice do not precisely match treatment guidelines. Therefore, researchers cannot rely on guidelines alone to identify treatment patterns and detect switches in therapy observed in RWD.

RESEARCH OBJECTIVES

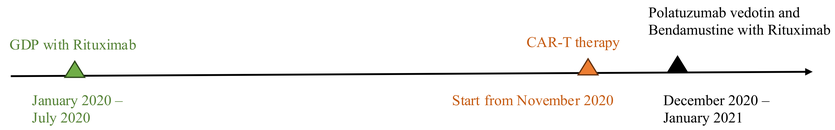
To develop an algorithmic approach to identifying switches in treatment lines for patients with DLBCL informed by clustering co-occurring chemo/immunotherapy drugs using unsupervised machine learning.

Exhibit 1. Examples of DLBCL Patient Treatment Patterns

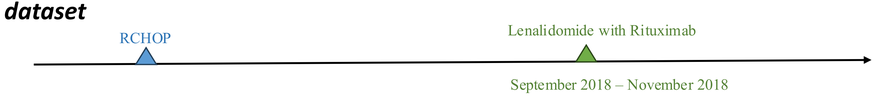


Example 1: CAR-T patient starting from their first line in the dataset

Example 2: CAR-T patient starting from their second line in the dataset



Example 3: Patient without CAR-T therapy starting from their first line in the dataset



RESULTS

Identifying Treatment Line Switches: Sets of line switch indicators were assessed for the ability to identify multiple lines of therapy among the group of 451 patients who actually received CAR-T and had sufficient pre-CAR-T treatment data available to assess (see Exhibit 3). Switches within each set were applied using a Boolean OR operator such that presence of any indicator in the set indicates that a switch has occurred. Patients receiving CAR-T therapy were presumed to have relapsed/refractory disease after at least two lines of therapy.

Final Sample of Eligible Patients: A set of switch indicators {A1, B1, C1, D1, E1, F1, G1, G2} was used to select patients who were presumably eligible for CAR-T therapy**.** To address potential left censoring (patients whose therapy was initiated before entering the TriNetX dataset), we included patients who started with a second-line therapy to the eligible group. 5,562 out of 13,289 DLBCL patients were identified as having at least two lines of treatment. 415 of 451 presumed eligible patients who received CAR-T were deemed eligible by the algorithm.

DISCUSSION

Some indicators were strong enough to indicate treatment line switches by themselves. However, requiring a strong indicator may lead to an over-specific result which may miss treatment line switches, excluding CAR-T eligible patients. We focused on characterizing various aspects with which a switch in line of treatment could be detected in order to select patients who were presumably eligible for CAR-T therapy. It is also essential to choose correct number of dimension and number of clusters for tradeoffs between sensitivity (the ability to identify eligible patients correctly) and specificity (the ability to rule out non-eligible patients correctly).

METHODS

Data: TriNetX Research Network electronic health records (2007-2022.04)

Sample: We selected 13,289 DLBCL patients who had DLBCL treatment records after October 18, 2017 and whose last DLBCL diagnosis code was at least 30 days after their initial diagnosis code. Among these DLBCL patients, 897 patients initiated CAR-T therapy before April 2022. 451 of 897 CAR-T patients had sufficient treatment records prior to CAR-T therapy to allow identification of prior systemic therapy.

Identifying Treatment Regimens: Among the 13,289 DLBCL patients, we found 270,043 unique drug records consistent with National Comprehensive Cancer Network (NCCN) and American Cancer Society (ACS) treatment guidelines and the literature. For each patient, DLBCL drugs delivered within a time period having no more than a 7-day gap were grouped into multi-drug treatment regimens.

Unsupervised Machine Learning: Multiple Correspondence Analysis (MCA) was used to identify dimensions comprising weighted combinations of co-occurring drugs in each regimen.[2] Regimens were then clustered into similar treatment groups by the mini-batch K-means algorithm,[3] using a subset of MCA dimensions explaining 95% of the underlying variance. The optimal number of clusters was chosen to minimize the Bayesian Information Criterion (BIC).

Treatment Line Switch Indicators: Multiple indicator dummy variables (A-G) were generated to identify likely beginning of a new line of treatment, including information on: whether or not a patient's regimen included a change in treatment cluster; time gaps between rounds of treatment; addition of new drugs (with special consideration of those recommend for use only in second-line or higher treatment), conditioning typically used prior to cellular therapy, for maintenance therapy or to manage toxicity (see Exhibit 2). Indicators A-E required changes in clusters and time gaps between treatment cycles larger than 14 days. A is the strongest with more restrictions, while E is the weakest indicator among A-E. Indicator F does not require cluster change or minimum of treatment time gap but focuses on addition of new drugs. Indicator G does not require cluster change but involves a long-time gap between treatments. A1-E1 indicators are stronger than A2-E2 indicators because the former set involves more restrictions (e.g. If a treatment regimen only showed for one time and did not repeat for the next time treatment, A1-E1 do not count it as a line switch but A2-E2 do.). F1 is stronger than F2 as F1 requires new drug(s) be the first time ever in a patient's records while F2 only compares drug changes with the previous treatment cycle. G1 is weaker than G2 as G2 requires longer time gap between treatment cycles.

Exhibit 2. Treatment Line Switch Indicators

Indicator	Cluster Change	Time Gap (Days)	Occur more than once consecutively	When previous treatment is Rituximab-only occuring once, it is not the first row	New drug first occurred	New drug added compare to previous treatment	Drug or treatment used only for second line
A1	Y	>14	Y	N	Y	Y	Y
B1	Y	>14	Y	N	Y	Y	N
C1	Y	>14	Y	N	N	Y	Y
D1	Y	>14	Y	N	N	Y	N
E1	Y	>14	Y	N	N	N	N
A2	Y	>14	N	N	Y	Y	Y
B2	Y	>14	N	N	Y	Y	N
C2	Y	>14	N	N	N	Y	Y
D2	Y	>14	N	N	N	Y	N
E2	Y	>14	N	N	N	N	N
F1	N	> 0	N	Y	Y	N	N
F2	N	> 0	Y	Y	N	Y	N
G1	N	> 60	N	N	N	N	N
G2	N	> 90	N	N	N	N	N

Exhibit 3. Distribution of Patients Identified with Sets of Switch Indicators

	A1/B1/C 1/D1/E1	A1/B1/C1 /D1/E1/G	A1/B1/C1/ D1/E1/G/F1	A/B/C/D/ E/G/F1	A/B/C/ D/E/G/F
# Patients with Line Switch(es)	2,096	3,280	3,799	4,160	5,147
# Patients Starting with 2nd Line	4,121	4,121	4,121	4,121	4,121
# Eligible Patients	4,859	5,439	5,562	5,718	6,587
# CAR-T Patients	411	416	419	420	420

Notes: (1) The number of eligible patients (row 3) included patients who started with a 2nd line therapy (row 2) to address potential left censoring; (2) A regimen coded as having any one from the set of indicators (e.g., A1 or B1 or C1 or D1 or E1 or F1 or any G) indicates a line switch has occurred.

CONCLUSION

Unsupervised machine learning is a promising approach for identifying treatment line switches. However, selection of switch indicators requires careful consideration of tradeoffs between sensitivity (including DLBCL patients actually eligible for CAR-T) and specificity (excluding patients who are not eligible).

REFERENCES

[1]Sermer, D., Batlevi, C., Palomba, M. L., Shah, G., Lin, R. J., Perales, M. A., ... & Sauter, C. (2020). Outcomes in patients with DLBCL treated with commercial CAR T cells compared with alternate therapies. *Blood advances*, *4*(19), 4669-4678.

[2]Le Phan, H. L., & Tortora, C. (2019). K-means clustering on multiple correspondence analysis coordinates. *Inst. Inf. Syst. Mark.*

[3]Violán, C., Roso-Llorach, A., Foguet-Boreu, Q., Guisado-Clavero, M., Pons-Vigués, M., Pujol-Ribera, E., & Valderas, J. M. (2018). Multimorbidity patterns with K-means nonhierarchical cluster analysis. *BMC family practice*, *19*, 1-11.

CONTACT

Yinan Wang, MPP
Pennsylvania State University
yuw415@psu.edu
Twitter: @YinanWangYW