OptumInsight Data

Notation	Variable definition & Characteristics
$i = 1, \ldots, N$	Index for the i-th patient; N = total number of patients = 14595
T_{i0}	Index VTE date
(T_{ij}, Y_{ij})	T_{ij} = the time of AC prescription; $T_{ij} \geq T_{i0} \forall j$; Y_{ij} = the prescribed anticoagulant at time T_{ij}, where $j = 1, \ldots, n_i$ with n_i being the number of anticogulant prescriptions after index VTE date, and $Y_{ij} \in A :=$ { DOACS (4 subcategories), LMWH, Warfarin, Other }
(Y_{i1}^*, Y_{i2}^*)	(Index AC, AC at 3 months) $Y_{i1}^* \in \{0, 1\}^K$, where $K = 7 = \|A\|$. Index AC is defined as the first AC after index VTE date; AC at 3 months is defined as the most recent AC prior to index VTE date + 90 days, and if a patient stopped on AC in the three months, it is recorded as "Not captured" (Warfarin + 60 days, INR + 42 days, LMWH/DOACS/Other + 30 days)
(T_{i1}^*, T_{i2}^*)	(Time of index AC, Time of AC at 3 months)
$\boldsymbol{V}_{ij}$	A vector of: copay, type of insurance, and provider information (TBD) at T_{ij}
$(T_{il}^{(L)}, \boldsymbol{X}_{il}^{(L)})$	$T_{il}^{(L)}$ = Time of lab tests $\boldsymbol{X}_{il}^{(L)}$ = A vector of: lab test type (hemoglobin, platelets, or GFR), and test result at time $T_{il}^{(L)}$, where $l = 1, \ldots, L_i$. Note that a majority of patients do not have lab test records.
D_{ij}	Days of supply of the AC at T_{ij}; $D_{ij} \in \{1, \ldots, D_{\max}\}$, where $D_{\max} = 90$.
S_{ij}	Dose of the AC at $T_{ij} = \frac{\text{quantity}}{\mathrm{D_{ij}}} \times$ strength; $S_{ij} \in \mathbb{R}_+$
$(T_{ir}^{(I)}, \boldsymbol{I}_{ir})$	$T_{ir}^{(I)}$ = time of the r-th INR test; I_{ir} = INR result at $T_{ir}^{(I)}$, where $r = 1, \ldots, R_i$.
$(T_{ik}^{(a)}, E_{ik}, \boldsymbol{C}_{ik}, P_{ik})$	$T_{ik}^{(a)}$ = the k-th admission date, $T_{ik}^{(a)} \geq T_{i0}$ E_{ik} = length of the k-th stay in days, $E_{ik} \geq 1$ $\boldsymbol{C}_{ik}$ = a binary vector indicating the ICD-9 codes associated with the admission; $\boldsymbol{C}_{ik} \in \{0, 1\}^M$ where M = the number of (tree-structured) ICD-9 codes $k = 1, \ldots, K_i$, where K_i = the number of admissions; P_{ik} = place of service (POS) (TBD)
$\boldsymbol{X}_i^{(1)}$	All time-variant covariates of the above, i.e. $\boldsymbol{V}_{ij}, j = 1, \ldots, n_i$; $(T_{il}^{(L)}, \boldsymbol{X}_{il}^{(L)}), l = 1, \ldots, L_i$; D_{ij}; S_{ij}; $(T_{ir}^{(I)}, \boldsymbol{I}_{ir}), r = 1, \ldots, R_i$; $(T_{ik}^{(a)}, E_{ik}, \boldsymbol{C}_{ik}, P_{ik}), k = 1, \ldots, K_i$.

$\boldsymbol{X}_i^{(2)}$	All time-invariant covariates of the i-th patient: 1. index VTE date 2. index cancer type 3. index cancer date 4. gender 5. SES: education, occupation, division, race, federal poverty level, home ownership, income range, networth range 6. indicator for having a surgery within 30 days prior to index VTE date 7. indicator for smoking within 30 days prior to index VTE date 8. place of service associated with index VTE date

Notes:

- Provider level information colored in blue is unidentifiable yet.
- Lab tests other than INR tests, i.e. hemoglobin, platelets, and GFR, can be either time-variant or time-invariant. If all records of such lab tests are considered, then they are denoted as $(T_{il}^{(L)}, \boldsymbol{X}_{il}^{(L)})$. If only the most recent lab tests within 30 days prior to index VTE date is considered, then they will go into the time-invariant covariate vector $\boldsymbol{X}_i$.
- Since ICD-9 codes are tree-structured, all diagnoses are tree-structured.

Scientific question:

1. Predicting the distribution of index AC and AC at 3 months given all covariates:
$$\left[Y_{i1}^*, Y_{i2}^* \middle| \boldsymbol{X}_i^{(1)}, \boldsymbol{X}_i^{(2)} \right],$$
where $[A|B]$ denotes the conditional distribution of A given B.
2. Predicting the anticoagulant prescription pattern after index VTE date:
$$\left[Y_{i1}, \ldots, Y_{in_i} \middle| \boldsymbol{X}_i^{(1)}, \boldsymbol{X}_i^{(2)} \right].$$

Features of the data:

1. Repeated multivariate outcomes: multiple drugs are prescribed repeatedly
2. Semi-regular time points: days of supply are commonly 30 days; less common are 15 days and 90 days, etc. Days of supply predict the time of the next prescription reasonably well.
3. Interrupted time points: information is always lost during hospitalization periods.

Random thoughts

- For predicting the anticoagulant pattern with covariates: Titsias, Michalis K., Christopher C. Holmes, and Christopher Yau. Statistical inference in hidden Markov models using k-segment constraints. Journal of the American Statistical Association 111, no. 513 (2016): 200-215.
- For predicting hospitalization associated with anticoagulant prescription patterns, we may need to look at a variety of reasons for hospitalization. These include medical diagnoses such as cancers, comorbidities, and other reasons.
- Consider a multi-category propensity score? $P(Y_{ij}|\boldsymbol{X}_{ij})$, where Y_{ij} = AC fill at time T_{ij}, and $\boldsymbol{X}_{ij}$ =covariate information up to time T_{ij}.