

# 1 **Machine learning enabled subgroup analysis with real-world** 2 **data to inform better clinical trial design**

3 Jie Xu<sup>1,\*</sup>, Hao Zhang<sup>1,\*</sup>, Hansi Zhang<sup>2</sup>, Jiang Bian<sup>2,#</sup>, Fei Wang<sup>1,#</sup>

4

5 <sup>1</sup>Department of Population Health Sciences. Weill Cornell Medical College. Cornell University.

6 <sup>2</sup>Department of Health Outcomes and Biomedical Informatics. College of Medicine. University of  
7 Florida.

8

9 <sup>\*</sup>Equal Contribution.

10

11 <sup>#</sup>Corresponding Authors.

12

13 Jiang Bian.

14 2197 Mowry Rd, Suite 122.

15 Gainesville. FL. 32610. USA.

16 [bianjiang@ufl.edu](mailto:bianjiang@ufl.edu)

17

18 Fei Wang.

19 425 East 61 Street.

20 New York. NY 10065. USA.

21 [few2001@med.cornell.edu](mailto:few2001@med.cornell.edu)

22

23 Number of words (main text): 4799

## 24 **Abstract**

25 Restrictive eligibility criteria for clinical trials may limit the generalizability of treatment  
26 effectiveness and safety to real-world patients. In this paper, we propose a machine learning  
27 approach to derive patient subgroups from real-world data (RWD), such that the patients within  
28 the same subgroup share similar clinical characteristics and safety outcomes. The effectiveness  
29 of our approach was validated on two existing clinical trials with the electronic health records  
30 (EHRs) from a large clinical research network. One is the donepezil trial for Alzheimer's disease  
31 (AD), and the other is the Bevacizumab trial on colon cancer (CRC). The results show that our  
32 proposed algorithm can identify patient subgroups with coherent clinical manifestations and  
33 similar risk levels of encountering severe adverse events (SAEs). We further exemplify that  
34 potential rules for describing the patient subgroups with less SAEs can be derived to inform the  
35 design of clinical trial eligibility criteria.

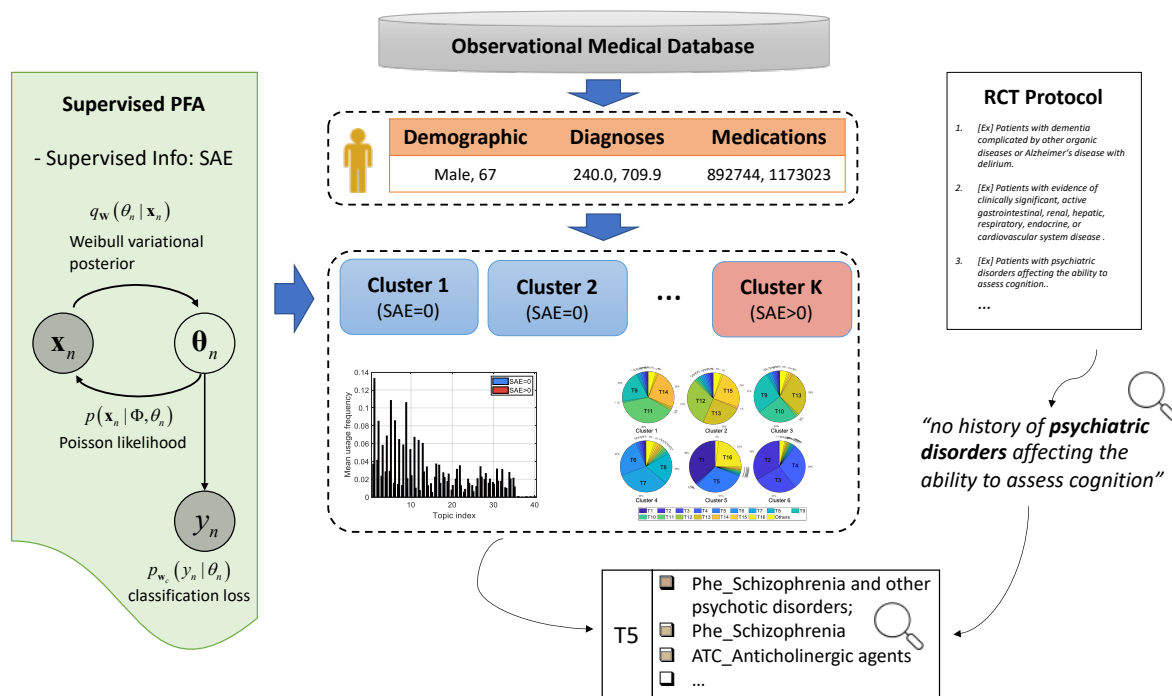
## 36 **Introduction**

37 Randomized controlled trials (RCTs) are the golden standard for investigating drug  
38 effectiveness (1). In designing an RCT, stringent eligibility criteria (EC) need to be applied to  
39 appropriately define a study population so that the drug effectiveness can be reliably and safely  
40 evaluated. In practice, trial designers often adopt eligibility criteria from existing similar trials,  
41 without too much consideration on their applicability and adaptability to the new trials. Moreover,  
42 many pivotal Phase III clinical trials continue to apply the same highly restrictive eligibility criteria  
43 as in the corresponding Phase I and Phase II trials. These oftentimes make the eligible trial  
44 participants not representative of the real-world patient population who will receive the treatment  
45 (2). For example, the elderly are often excluded from cancer and Alzheimer's disease drug  
46 development trials and are therefore under-represented among the primary target populations  
47 for these drugs (3, 4).

48 Although excessive or overly restrictive EC may lower the risk of the study populations for  
49 encountering adverse events (5-7), they usually lead to low population representativeness (thus,  
50 low trial generalizability), and subsequently, treatment effectiveness could be reduced, and the  
51 likelihood of adverse outcomes could increase when the treatment entered real-world clinical  
52 practice (8). Therefore, as recommended by regulatory agencies such as the U.S. Food and  
53 Drug Administration (FDA), broadening eligibility criteria during enrollment to increase the  
54 diversity of the clinical trial population and thus improve generalizability is an important trend (9).

55 Trial generalizability is largely dependent on the representativeness of the study population with  
56 respect to target population to which the study results are intended to be applied (8). In recent  
57 years, the rapid adoption of electronic health records (EHR) systems has led to large integrated  
58 clinical data warehouses and interoperable data networks, which made it possible for  
59 accumulating large amounts of real-world patient data. Such data provide us a unique  
60 opportunity for simulating the study and target populations of clinical trials. The goal of this study  
61 is to develop machine learning approaches for mining insights from RWD that could be  
62 informative of clinical trial eligibility criteria design. In particular, to account for the heterogeneity

63 of the real-world population, we introduce a novel transparent and outcome-guided probabilistic  
 64 model to identify the subphenotypes who took a specific medication (Fig. 1). The patients within  
 65 the same subphenotype do not just share similar clinical characteristics, but also face a similar  
 66 risk of encountering severe adverse events (SAEs) after they take the drug. We assume certain  
 67 compositions (co-occurrence patterns) of the clinical events within an individual’s EHR can  
 68 determine those subphenotypes and propose a novel weakly supervised clinical topic modeling  
 69 approach to identify those subphenotypes. Here each clinical topic represents a certain clinical  
 70 event composition pattern learned from the patient EHRs.



71  
 72 **Fig. 1 Model overview.** Clinical events including demographics, diagnoses, and medications  
 73 were extracted from RWD to represent patients. Supervised Poisson factor analysis (PFA) is  
 74 applied to identify patient subgroups with coherent clinical latent topics and outcomes measured  
 75 by SAEs. Subgroups with less SAEs can be derived to inform the design of clinical trial eligibility  
 76 criteria.

77 We evaluated our algorithm with two real-world clinical trials: (1) a hallmark randomized Phase  
 78 III trial (i.e., NCT00478205) that compares the effects of 23 mg to 10 mg Donepezil in treating  
 79 patients with Alzheimer’s disease that led to the FDA approval of the 23 mg Donepezil; and (2)  
 80 another randomized Phase III trial studying two different combination chemotherapy regimens  
 81 with or without bevacizumab (trade name Avastin) in high-risk stage II/III colon cancer patients.  
 82 We used RWD data from the OneFlorida Clinical Research Consortium -- a statewide clinical  
 83 data warehouse containing large collections of linked EHR, administrative claims, vital statistics,  
 84 and cancer registry data among others, covering more than 15 million (~60%) Floridians (10).  
 85 Our results showed that, for both trials, 1) a significant number of real-world patients who  
 86 satisfied the ECs and took the drug suffered from SAEs; 2) our model can clearly identify the  
 87 patient subgroups who are more likely to suffer or not suffer from SAEs as subphenotypes in a  
 88 transparent and interpretable way. We have also inferred the clinical topics from the

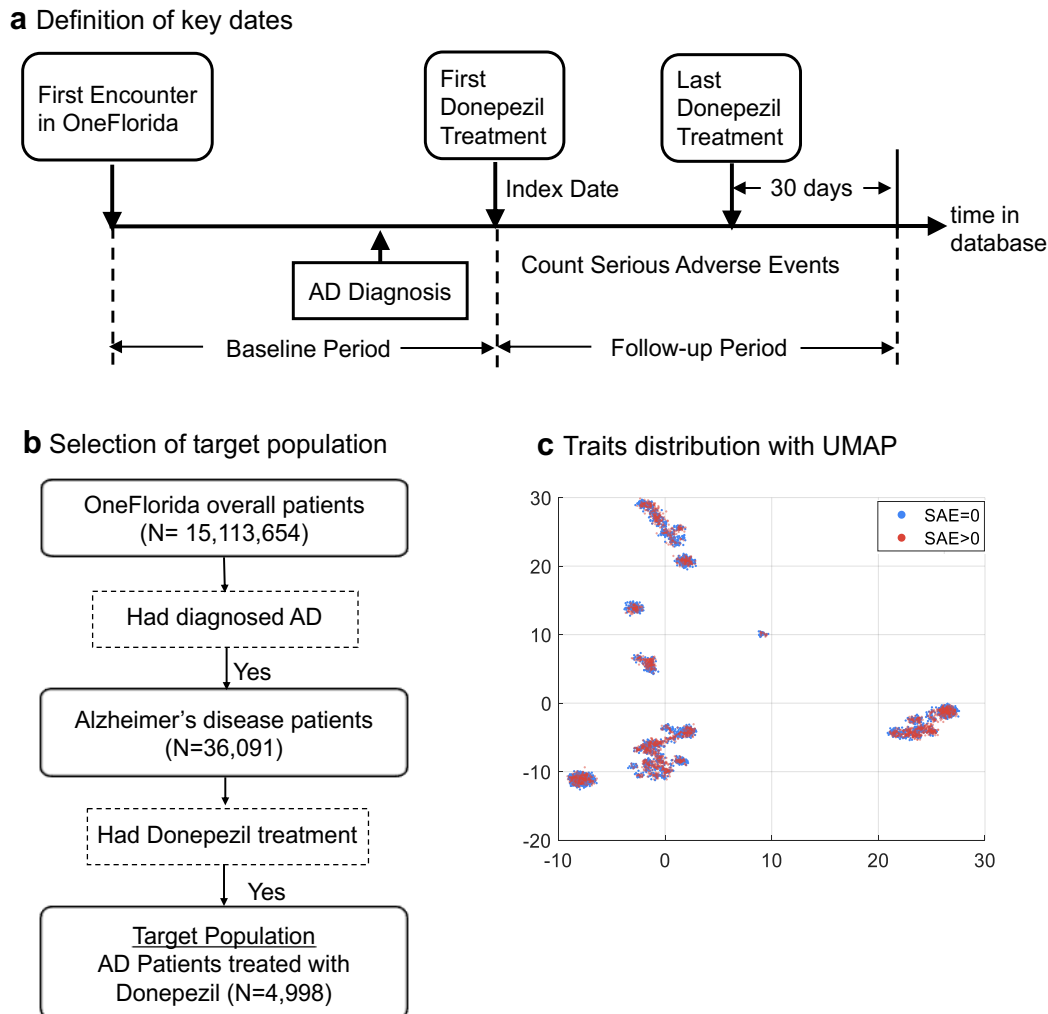
89 subphenotypes with or without SAEs, which reveal meaningful combinations of clinical features  
90 across multiple data types and provide data-driven recommendations for refining the ECs of  
91 clinical trials.

## 92 **Results**

93  
94 **Target population of the donepezil clinical trial.** NCT00478205 is a double-blind, double-  
95 dummy trial that compares different dosages (23 mg vs. 10 mg) of donepezil for treating  
96 patients with moderate to severe Alzheimer's disease (11). The real-world individual-level EHR  
97 data are from the OneFlorida Clinical Research Consortium. We constrained the target  
98 population as those who (1) were diagnosed with AD, and (2) treated with donepezil. A total of  
99 4,998 unique patients (mean (SD) age, 77.53 (9.9) years) were identified from OneFlorida (Fig.  
100 2b, Table 1).

101 To determine whether a patient experienced SAEs related to donepezil, we first compiled the list  
102 of SAEs reported in the original trial in ClinicalTrials.gov, and then identified relevant diagnosis  
103 codes (i.e., International Classification of Diseases, Ninth/Tenth Revision, Clinical Modification  
104 [ICD-9/10-CM]) for each SAE. Based on the severity grading scale defined in the Common  
105 Terminology Criteria for Adverse Events (CTCAE) (12), adverse events leading to  
106 hospitalization or mortality are grade 3 and above are considered as SAEs in our study. Further,  
107 we defined that the SAE diagnosis codes have to occur 1) after the first donepezil treatment,  
108 and 2) before 30 days after the last donepezil treatment (Fig. 2a). Among the patients in our  
109 target population, 3,063 (61.3%) had no SAE while 1,935 (38.7%) had at least one SAE. We  
110 dropped the ECs which are not computable (e.g., subjective eligibility criteria such as "written  
111 informed consent" is not computable, but also not going to affect our analysis, since we can  
112 safely assume all patients treated with donepezil in the real-world are consented) and extracted  
113 study traits corresponding to each computable EC from the OneFlorida data (Table 1).

114 We first represent each patient using the extracted traits as a vector and check whether patients  
115 with and without SAEs can be well separated. Fig. 2c shows the 2D embeddings of patient traits  
116 with Uniform Manifold Approximation and Projection (UMAP) (13). We color each sample  
117 based on whether the patient had SAEs or not. As shown in the figure, the two clusters, patients  
118 with (#SAE>0) vs. without (#SAE=0) are mixed up with each other, which indicates that the trial  
119 ECs cannot guarantee the safety for real-world patients. In addition, we examined the  
120 differences of the study traits value distributions between the two groups through Chi-square  
121 tests and summarized the results in Table 1, from which we observe that many traits were not  
122 significantly different (statistical sense, with  $p>0.05$ ) including memantine ( $p=0.145$ ),  
123 antidepressant ( $p=0.590$ ), basal/squamous cell carcinoma of the skin ( $p=0.275$ ), galantamine  
124 ( $p=0.190$ ), severe lactose intolerance ( $p=0.219$ ), and clinically significant Hepatic ( $p=0.105$ ).  
125 This suggests that the trial ECs can further be optimized with real-world evidence.  
126



127

128 **Fig. 2 The donepezil trial for Alzheimer's disease.** **a** Definition of key dates; **b** Selection of  
 129 target population; **c** Traits distribution with UMAP among two patients' groups with or without  
 130 SAE.

131 **Study design of the donepezil clinical trial.** The beginning of the treatment is termed the  
 132 index date. We set the index date in our donepezil trial to the first (ever) observed prescription  
 133 date of donepezil. We refer to the observed time before the index date as the baseline period  
 134 and use the information (i.e., demographic, diagnosis and medication history) collected during  
 135 that time period for analysis. The period from the index date to the last donepezil prescription  
 136 plus 30 days was set as the follow-up period, from which the SAE information is collected. Fig.  
 137 2a illustrates the key dates in our study design.

138 We applied supervised Poisson factor analysis (SPFA) to the collected information and used the  
 139 occurrence of SAE as the supervision to guide the learning process of PFA. Similar to other  
 140 topic modeling approaches (14), SPFA first compressed the clinical events into a set of  
 141 overlapping groups (i.e., topics). Novel patient representations are derived on top of these  
 142 topics (which are groups of clinical events that tend to co-appear in the same visit within the  
 143 RWD). K-means clustering is then performed on these new patient representations to identify

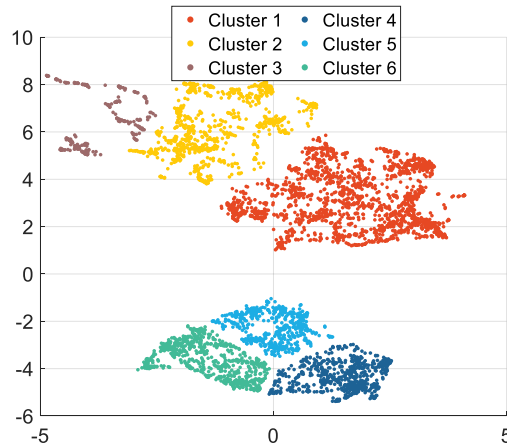
144 the clusters as subphenotypes. To choose the optimal number of topics K, we used all samples  
145 to learn the supervised topic model and then evaluated the topic coherence by normalized  
146 pointwise mutual information (NPMI) value (15), and the classification performance by ROC-  
147 AUC. We set K=40 for subsequent analyses as it achieved the highest ROC-AUC with large  
148 topic coherence values (Supplementary Fig. 1). The model robustness with respect to K will be  
149 discussed in the methods.

150 As discussed in (16), according to the silhouette score (17), we select the most appropriate  
151 number of clusters that provides the largest silhouette score. As a result, six clusters were  
152 derived with K-means for patients who had been diagnosed with AD and treated with Donepezil  
153 (Fig. 3a). Among the six clusters, two patient subgroups emerged: (1) the SAE group (#SAE>0)  
154 containing clusters 4, 5, and 6, and (2) the non-SAE group (#SAE=0) including clusters 1, 2 and  
155 3. As shown in Fig. 4a, the two patient subgroups (i.e., #SAE=0 vs. #SAE>0) are well separated.  
156 Specifically, for the SAE group, 1,915 out of total 1,935 patients (99.0%) encountered SAE  
157 events; for non-SAE group, 3,014 out of 3,063 patients (98.4%) did not have any SAE events.

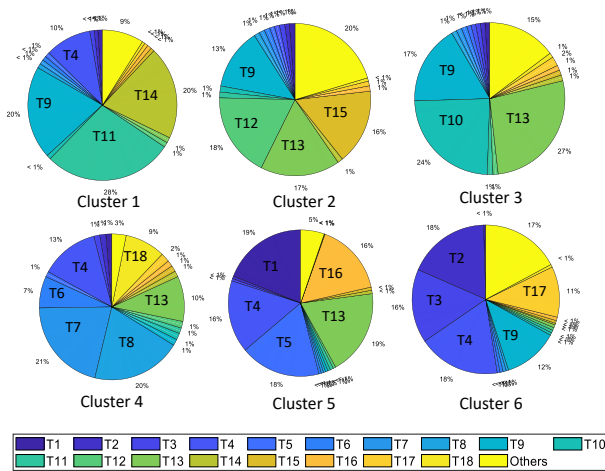
158  
159 We examined the distribution of the 40 topics across the two subgroups (Fig 4b). According to i)  
160 the mean topic weights (MTW) over all samples (evaluation for the usage of each topic on the  
161 cohort) and ii) Mann–Whitney U (MWU) test (18) on MTW of SAE subgroup (#SAE>0) versus  
162 non-SAE subgroup (#SAE=0) (evaluation for difference of topic usage between two groups), 18  
163 typical topics were selected to be analyzed detailly. Specifically, 15 topics (except for T4, T9,  
164 and T13 in Fig 2c) whose MWU p-value of SAE subgroup versus non-SAE subgroup are  
165 smaller than 0.05 (significant difference between two groups) while MTW on all samples are  
166 larger than 0.1 were selected. Of the 15 typical significant-difference topics, 10 topics (T1~T3,  
167 T5~T8, T16~T18, denoted by red in Fig. 3c) were characterized as likely to align with the SAE  
168 subgroup (#SAE>0) and the other 5 topics (T10~T12, T14~T15, denoted by blue in Fig. 3c)  
169 align with the other non-SAE subgroup (#SAE=0). Besides, 3 topics (T4, T9, and T13) whose p-  
170 values are larger than 0.05 but MTW on all samples are the top-three largest ones were also  
171 selected. These 3 topics (T4, T9, and T13) are shared by all clusters.

172

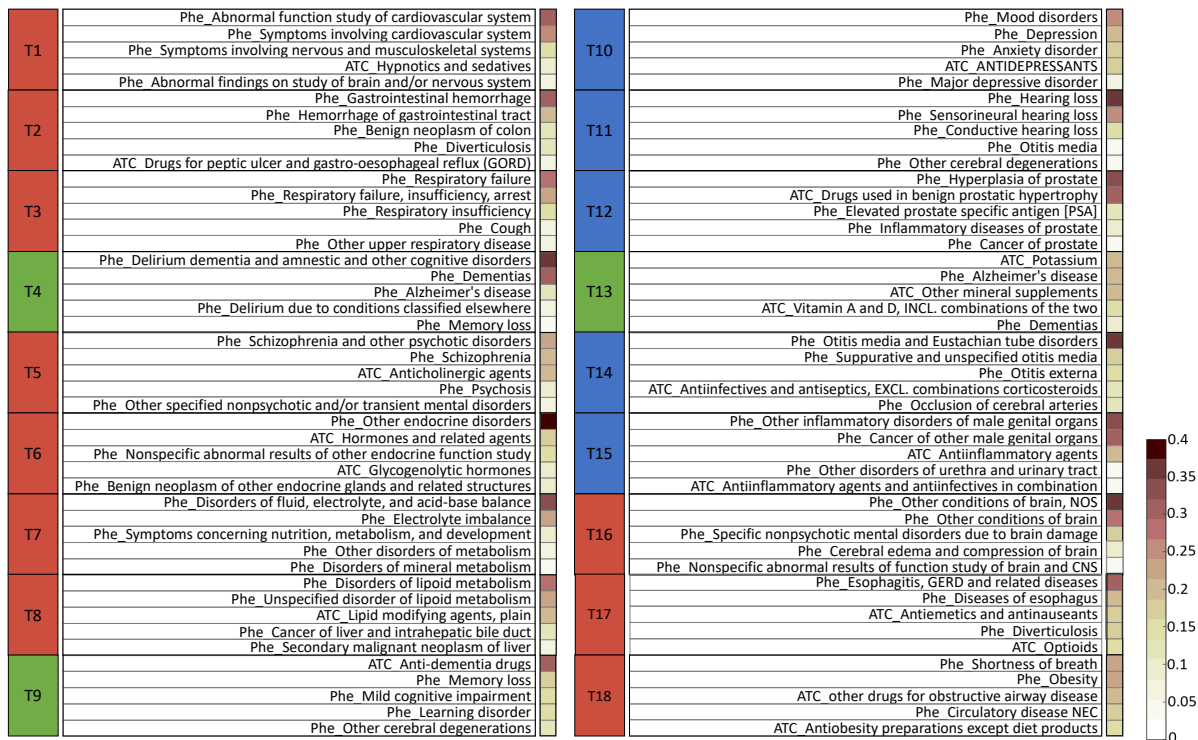
**a** Data visualization with UMAP



**b** Topic distributions of each cluster

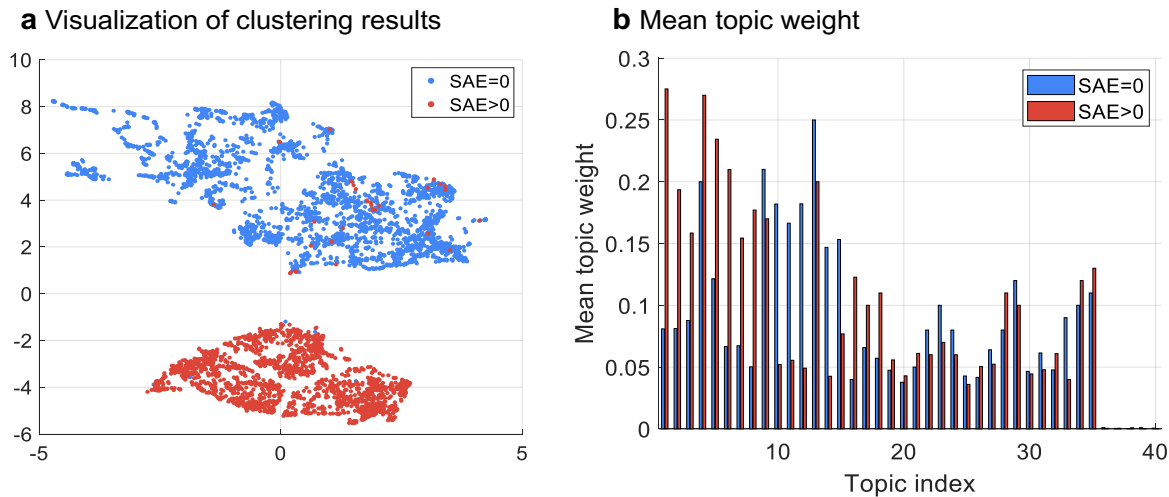


**c** Top clinical features from certain disease topics



173  
174  
175  
176  
177  
178  
179

**Fig. 3 Clustering results of AD target population.** **a** Data visualization with UMAP; **b** Topic distributions of each cluster; **c** Top clinical features from certain disease topics, where the red, blue, and green topics represent the commonly used topics for clusters 4, 5, 6 (most of samples are with SAE>0), clusters 1, 2, 3 (most of samples are with SAE=0), and all clusters, respectively.



180  
 181 **Fig. 4 Donepezil clinical trial. a** Visualization of clustering results; **b** Mean topic weight (MTW)  
 182 of all topics on two groups, where x-axis is the topic index and y-axis is the MTW of each topic  
 183 on two subgroups.  
 184

185 We then examined the relevance of our learned disease topics by qualitatively assessing the  
 186 coherence of the five most prevalent clinical events (i.e., diagnosis and medication codes) for  
 187 each topic (19). We interpreted these 18 topics based on these clinical events and found that  
 188 many of them were specific to different diseases (Fig 3c). These 18 topics can be divided into  
 189 three categories. Specifically, i) T4, T9, and T13 includes dementia, memory loss and cognitive  
 190 impairment related events, which are commonly used among all clusters. These topics  
 191 represent the common diseases and drugs in the cohort. ii) T1 is related to cardiovascular  
 192 diseases. T2 is related to gastrointestinal diseases. T3 is about respiratory disorders. T5 is  
 193 related to psychotic disorders, especially Schizophrenia and relevant treatments (anticholinergic  
 194 agents) (20). T6 is related to endocrine disorders. T7 is about metabolism disorders such as  
 195 mineral metabolism disorder. T8 includes lipid metabolism and secondary malignant neoplasm  
 196 or cancer of the liver. There have been prior studies showing the relationship between these two  
 197 types of diseases (21). T16 includes various conditions or disorders of brains, which is closely  
 198 related to AD. T17 talks about the diseases and treatments of esophagus such as  
 199 gastroesophageal reflux disease (GERD). T18 is about obesity and some related complications  
 200 and drugs. Six clusters were derived which is characterized by some related clinical topics:  
 201 cluster 1 (N=1811; 36.23%), patients with disorders of ears or eyes (T11 and T14); cluster 2  
 202 (N=939; 18.79%), patients with diseases of urinary system (T12 and T15); cluster 3 (N=331;  
 203 6.62%), patients with depression or mood disorder (T10 and T13); cluster 4 (N=667; 13.35%),  
 204 patients with disorders of endocrine and metabolism (T6, T7, and T8); cluster 5 (N=548;  
 205 10.96%), patients with different diseases of brain (T1, T5, and T16); cluster 6 (N=702; 14.05%),  
 206 patients with diseases of digestive and respiratory systems (T2, T3, and T17).

207 The topics, commonly used in the SAE subgroup, may have close relation with exclusion criteria  
 208 in EC. T11 and T14 are related to diseases of ears such as hearing loss in T11 and otitis media  
 209 in T14 with related drugs, which have some shared characteristics with AD (22). T12 and T15  
 210 related to various diseases are about male genital organs such as hyperplasia of prostate and



211 urinary tract disorder. T10 is associated with mood disorders and depression, where their  
212 relationship to dementia and AD is unclear (23). These topics are mostly related to relatively  
213 mild conditions with no direct connection to the diagnosis of AD. In other words, most  
214 phenotypes associated with the non-SAE subgroup are mild aging-related chronic comorbidities  
215 that are common in older adults and do not lead to serious adverse effects. Therefore, diseases  
216 which are more associated with the non-SAE patient subgroup may have a lower probability of  
217 causing adverse serious events.

218 We further analyzed the association between the inferred topics with the SAE subgroup and the  
219 extracted computable eligibility criteria. We will explain why our model is more powerful in  
220 distinguishing two subgroups compared to traits.

221 **Align with exclusion criterion in clinical trials.** Firstly, we found that topics (T1~T8) aligned  
222 with the SAE patient subgroup (#SAE>0) are highly associated with the exclusion criteria of the  
223 donepezil trial. Here, [Ex] represents the “exclusion criterion” of clinical trials.

224 • “[Ex] Patients with evidence of clinically significant, active gastrointestinal, renal, hepatic,  
225 respiratory, endocrine, or cardiovascular system disease (including history of life-  
226 threatening arrhythmias).”

227 We found T1 (cardiovascular), T2 (gastrointestinal), T3 (respiratory), T6 (endocrine) are  
228 highly related to this exclusion criterion. Disorders of lipid metabolism (i.e., T8) is a  
229 major contributing factor that mediates the development of cardiovascular diseases,  
230 especially atherosclerosis (24).

231 • “[Ex] Patients with dementia complicated by other organic diseases or Alzheimer's  
232 disease with delirium.”

233 We found T4 (delirium) is related to this criterion. Although T4 is a topic shared by non-  
234 SAE subgroup (cluster 1) and SAE subgroup (clusters 4, 5, 6), its proportion and MTW  
235 are higher in SAE group (Fig 3b and 4b).

236 • “[Ex] Patients with psychiatric disorders affecting the ability to assess cognition such as  
237 schizophrenia, bipolar or unipolar depression. Patients with clinically significant sleep  
238 disorders will also be excluded unless these are controlled by treatment and clinically  
239 stable for > 3 months prior to screening.”

240 We found T5 (psychotic disorders) is related to this criterion.

241 • “[Ex] Patients with any conditions affecting absorption, distribution, or metabolism of the  
242 study medication (e.g., inflammatory bowel disease, gastric or duodenal ulcers, hepatic  
243 disease, or severe lactose intolerance).”

244 We found T7 is related to this criterion. T7 is characterized as other disorders of  
245 metabolism include disorders of fluid, electrolyte, and acid-base balance. Acid-base and  
246 electrolyte homeostasis is essential for the proper functioning of many metabolic  
247 processes and organ functions in the body (25). Phenotypes in T7 seem to have effects  
248 on absorption or metabolism of the study medication.

249 • “[Ex] Patients with a history of cancer (does not include basal or squamous cell  
250 carcinoma of the skin) treated within 5 years prior to study entry, or current evidence of

251 *malignant neoplasm, recurrent, metastatic disease. Males with localized prostate cancer*  
252 *requiring no treatment would not be excluded.”*

253 Although we annotated T8 to the disorder of lipid metabolism, it also includes  
254 secondary malignant neoplasm or cancer of the liver. As the disorder of lipid metabolism  
255 plays an important role in carcinogenesis and development, we concluded T8 is also  
256 related to this criterion.

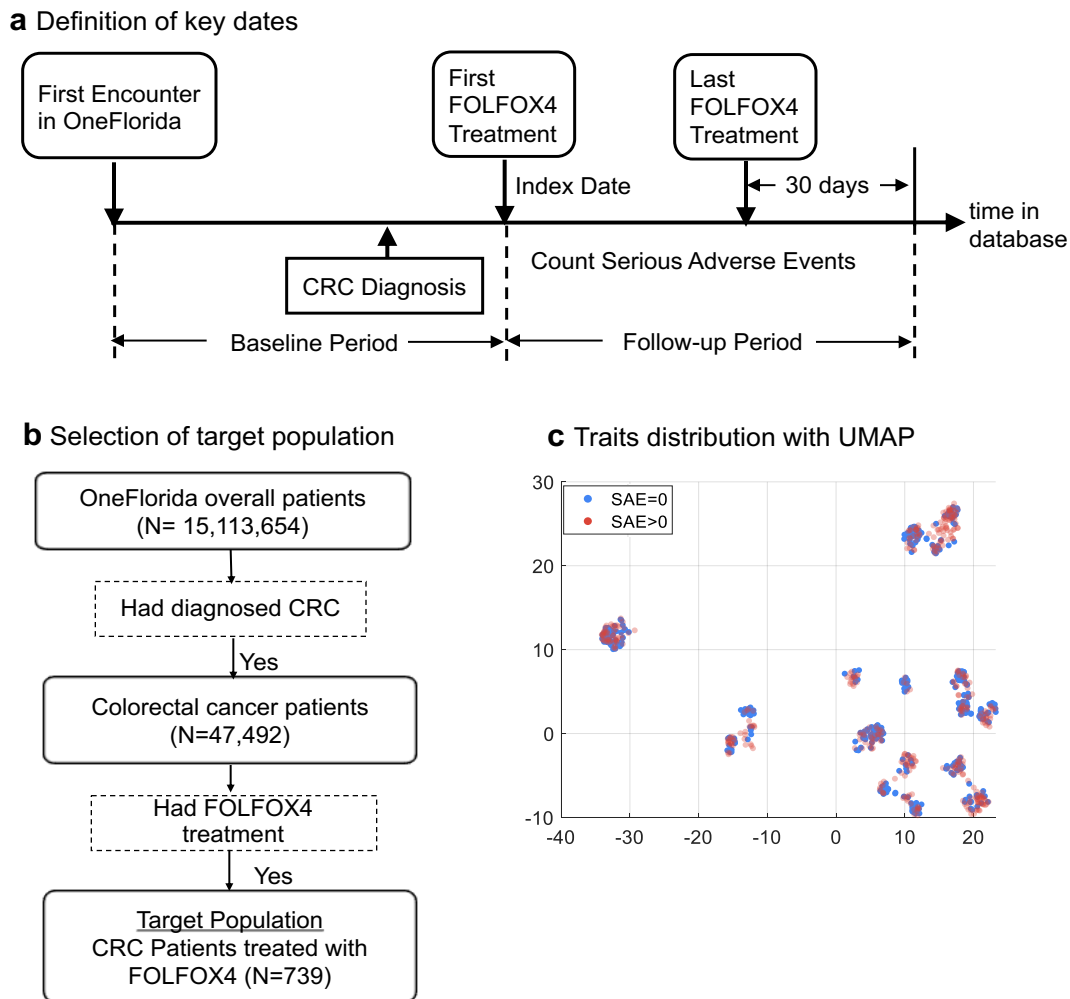
257 **Different from and move beyond exclusion criteria in clinical trials.** Secondly, we found  
258 that some topics (T16~T18) have no direct connections with exclusion criteria. These topics  
259 describe some diseases, such as disorders of the brain in T16, diseases of esophagus in T17,  
260 and obesity in T18, which may have a severe impact for older people. For example, in (26),  
261 authors discussed associations of obesity with numerous life-threatening diseases. In (27),  
262 authors discussed the potential life-threatening gastroesophageal reflux disease (GERD). In  
263 other words, our method may discover some diseases that have potential outcomes of serious  
264 adverse events for elder patients, but are ignored by the original EC. This is the first reason why  
265 our method, learning knowledge from more abundant features, can separate SAE and non-SAE  
266 subgroups well compared with original traits from EC. Furthermore, even for one disease that  
267 appears in both SAE-associated topics and EC, more abundant features provide more detailed  
268 insights. For example, for gastrointestinal disease, EC only said “[Ex] *Patients with evidence of*  
269 *clinically significant active gastrointestinal disease*”, which may be a coarse description.  
270 However, the learned topics, T2 and T17, discover more detailed diseases or drugs about  
271 gastrointestinal disease.

272 There are also eligibility criteria that have no apparent relations with these inferred topics likely  
273 due to two main reasons: (1) since we did not use procedure codes and laboratory tests in the  
274 topic modeling process, eligibility criteria related to procedures and lab tests were missing in  
275 these inferred topics; and (2) there are eligibility criteria that are not used to ensure patient  
276 safety but to ensure treatment efficacy.

277 **Target population of the bevacizumab (Avastin) clinical trial.** To demonstrate the  
278 generalizability of our approach, we analyzed another RCT (NCT00112918) which studied a  
279 combination chemotherapy with or without bevacizumab in treating patients who have  
280 undergone surgery for high-risk stage II or stage III colon cancer. Specifically, the trial studied  
281 three combination chemotherapies: (1) the control arm, a standard-of-care (SoC) treatment,  
282 FOLFOX4 (i.e., intermittent fluorouracil/leucovorin with oxaliplatin), (2) FOLFOX4 plus  
283 bevacizumab, and (3) XELOX (i.e., capecitabine with oxaliplatin) plus bevacizumab. In our  
284 RWD analysis, we set the target population as patients who (1) were diagnosed with colorectal  
285 cancer (CRC), and (2) treated with FOLFOX4 (Fig. 5(b), Table 2). Since FOLFOX4 was the SoC  
286 treatment, patients who were treated with FOLFOX4 in the real-world could potentially benefit  
287 from the new combination chemotherapies investigated in this trial. A total of 739 patients  
288 (mean age, 57.49 with a standard deviation of 11.2 years old) were identified (out of a total of  
289 47,492 CRC patients) from OneFlorida.

290 Similar to the Donepezil trial, we defined the index date as the date of first FOLFOX4 treatment  
291 after CRC diagnosis. The baseline period was defined as the time period before the index date,

292 and the follow-up period was defined as the time period between index date and the last  
 293 FOLFOX4 treatment date plus 30 days. SAE events were collected from the follow-up period.  
 294 This study design is depicted in Fig. 4(a). Among all CRC patients who received FOLFOX4 after  
 295 diagnosis, 347 (47.0%) had no SAE, while 392 (53.0%) had at least one SAE. We examined the  
 296 computability of each eligibility criterion and constructed queries to extract study traits  
 297 corresponding to the computable ones based on the OneFlorida data. The identified traits  
 298 included patient demographics (e.g., age) and medical history (e.g., medication, treatment). We  
 299 conducted Chi-square tests on the two patient subgroups, i.e., patients had SAEs (#SAE>0) vs  
 300 patients who did not (#SAE=0). We found that the p values of most study traits are larger than  
 301 0.05, except for metastatic disease (p=0.026), parenteral anticoagulants (p<0.001), myocardial  
 302 infarction (p<0.001) and significant and thrombolytic agent (p=0.003). The p value of the trait  
 303 related to “inability to take oral medication” was smaller than 0.05 after adjusting for  
 304 demographics (Table 2).

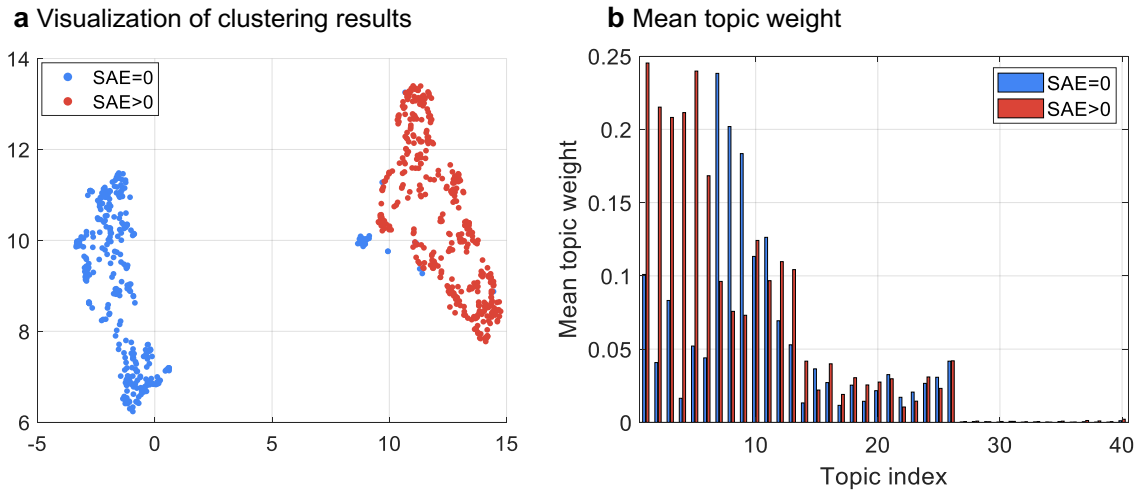


305  
 306  
 307 **Fig. 5 The Bevacizumab (Avastin) clinical trial for colon cancer.** **a** Definition of key dates; **b**  
 308 Selection of the target population; **c** Traits distribution with UMAP among the two patients’  
 309 groups with or without SAEs.  
 310

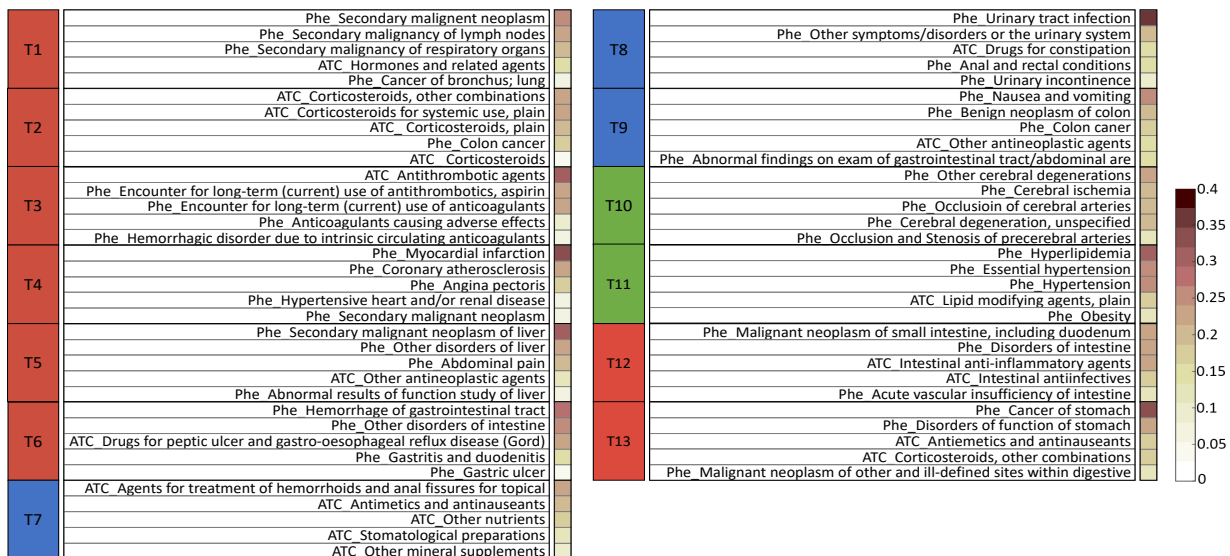
311 Similar to the donepezil trial, we applied SPFA to the CRC target population. The patient  
312 features we collected from the baseline period include demographics, diagnosis (as phecodes)  
313 and medications (as ATC classes)). Similar to what we have done in analysis of Donepezil trial,  
314 we evaluated the topic coherence by NPMI value (15), and the classification performance by  
315 ROC-AUC to decide the number of topics. We set the number of topics  $K=40$  since it achieves  
316 higher ROC-AUC and NPMI (Supplementary Fig. 1). Fig. 6a shows the UMAP embeddings of  
317 new patient representations induced by SPFA. From the figure we can observe two well-  
318 separated patient subgroups, which can be identified by K-means clustering. One group (red) is  
319 mostly associated with SAE (317 of 347 patients, or 91.4% of the patients in this subgroup  
320 encountered SAE) and the other group (blue) is free of SAE (393 patients). We checked the  
321 patient group proportions for 40 topics in the two clusters (Fig. 6b). Of the 40 topics in the model,  
322 patients' proportion in 9 topics were larger than 0.075 and visually very different in two patient  
323 subgroups. Of these 9 typical topics, 6 topics (T1~T6) were characterized as likely to align with  
324 the SAE subgroup ( $\#SAE>0$ ) and the other 3 topics (T7~T9) align with the non-SAE subgroup  
325 ( $\#SAE=0$ ). We annotated the 9 topics based on the top EHR codes and found that most of the  
326 inferred topics were specific to different diseases or treatments (Fig.7, Supplementary Fig. 3).

327 Among all 40 topics, similar to the topic selection method for donepezil trial, we selected 13  
328 topics for detailed analysis (Fig. 7). According to the MTW on two groups (Fig. 6b,  
329 Supplementary Table 2), these topics can be divided into three categories: i) T1~T6 and  
330 T12~T13, represented as red, are associated with SAE subgroup; ii) T7~T9, represented as  
331 blue, aligned with the non-SAE subgroup, contain relatively mild diseases and were not directly  
332 related to the diagnosis of colon cancer; iii) T10~T11, represented as green, are often shared  
333 on two subgroups. Specifically, T1 is annotated to the use of corticosteroids, with the three of  
334 the top five codes being specific corticosteroids treatments. T2 is related to antithrombotic  
335 agents. T3 is about malignant neoplasm, where the first three main codes are all correlated with  
336 secondary malignant neoplasm and one code is about cancer, and one code is hormones and  
337 related preparations drug class which is used to treat cancer. T4 is related to clinically  
338 significant (i.e., active) cardiovascular disease. T5 is also related to cancer, but more specific to  
339 the liver. Phenotypes in T6 include various kinds of disorders related to intestine, e.g., drugs for  
340 peptic ulcer and gastro-oesophageal reflux disease (GORD), hemorrhage of gastrointestinal  
341 tract, gastritis and duodenitis and gastric ulcer. T7 includes some commonly used drugs. T8  
342 talks about the disease and treatments about urinary tract infection, which is extremely common  
343 in the elderly (28). T9 is associated with gastrointestinal diseases such as nausea and vomiting.  
344 T10 and T11 are some common diseases such as or hyperlipidemia or hypertension. T12  
345 includes different diseases or treatments about the intestine. For example, authors in (29)  
346 discussed the relations between CRC and chronic intestinal inflammation. T13 is about the  
347 cancer or malignant neoplasms of digestive organs. Next, we analyze the connections of SAE-  
348 group related topics (the red one) with the EC in trial.

349  
350  
351



352  
 353 **Fig. 6 Bevacizumab (Avastin) clinical trial. a** Visualization of clustering results; **b** Mean topic  
 354 weight for 40 topics where x-axis is the topic and y-axis is the proportion of each group in that  
 355 topic.  
 356



357  
 358 **Fig. 7 Top clinical features from disease topics in Bevacizumab (Avastin) clinical trial.**

359 **Association analysis between topics and the eligibility criteria of bevacizumab (Avastin)**  
 360 **trial.** Firstly, we analyzed the association between the inferred topics and extracted eligibility  
 361 criteria. We compared each typical topic with the criteria. We found that topics, T1~T6, aligned  
 362 with the SAE subgroup (#SAE>0) are highly associated with some of the exclusion criteria in the  
 363 original trial.

- 364 • “[Ex] Chronic treatment with corticosteroids (dose of  $\geq 10$  mg/day methylprednisolone  
 365 equivalent) (excluding inhaled steroids).”  
 366 We found T1 (corticosteroids) is highly related to this exclusion criterion.

367 • “[Ex] Current or recent (within 10 days prior to study treatment start) use of full-dose oral  
368 or parenteral anticoagulants or thrombolytic agents for therapeutic purposes.”  
369 We found T2 (antithrombotic) is related to this criterion.

370 • “[Ex] Macroscopic or microscopic evidence of remaining tumour. Patients should never  
371 have had any evidence of metastatic disease (including presence of tumour cells in the  
372 ascites). The isolated finding of cytokeratin positive cells in bone marrow is not  
373 considered evidence of metastatic disease for purposes of this study.

374 [Ex] Other malignancies within the last 5 years (other than curatively treated basal cell  
375 carcinoma of the skin and/or in situ carcinoma of the cervix).

376 [Ex] Previous anti-angiogenic treatment for any malignancy; cytotoxic chemotherapy,  
377 radiotherapy or immunotherapy for colon cancer.”

378 We found T3 (malignant neoplasm) and T5 (disorders of liver) are related to these  
379 exclusion criteria. Although we annotated T5 as the disorder of liver, it also included  
380 malignant neoplasm of liver or antineoplastic agents.

381 • “[Ex] Clinically significant (i.e., active) cardiovascular disease. This includes, but is not  
382 limited to, the following examples: cerebrovascular accidents ( $\leq 6$  months prior to  
383 randomization), myocardial infarction ( $\leq 1$  year prior to randomization) ...”

384 We found T4 (cardiovascular) is characterized by myocardial infarction, coronary  
385 atherosclerosis, angina pectoris, which is very related to this criterion.

386 • “[Ex] Lack of physical integrity of the upper gastro-intestinal tract, malabsorption  
387 syndrome, or inability to take oral medication.”

388 We found T6 is related to this criterion.

389 Secondly, we analyzed the difference of learned topics with EC and discussed why our methods  
390 can separate SAE and non-SAE well. Some topics, T12 and T13, are beyond the exclusion  
391 criteria in trial, which provide more discriminant features to separate two subgroups. For  
392 instance, T12 is about intestinal diseases, which may bring a high mortality rate and often  
393 related to multiple comorbidities (30). Even for the topics that have relation with exclusion  
394 criteria as discussed above, the learned clinical topics provide more detailed information. For  
395 example, EC provides a rough description about corticosteroids as “Current or recent (within 10  
396 days prior to study treatment start) use of full-dose oral or parenteral anticoagulants or  
397 thrombolytic agents for therapeutic purposes”. Compared with it, the topic T2 contains more  
398 detailed drugs about corticosteroids. Therefore, these two benefits make our method provide a  
399 better separation for two subgroups.

400 Similar to the findings in the donepezil trial analysis, there are some criteria which have no  
401 relations with the learned topics. Some of these criteria have potential to be adjusted. For  
402 example, for the exclusion criterion “History or evidence upon physical examination of central  
403 nervous disease (CNS) disease (e.g., primary brain tumour, seizure not controlled with standard  
404 medical therapy, any brain metastases),” Exclusion of patients with brain metastases (BMs) in  
405 clinical trials is common. Although life expectancy may be reduced for some patients with BMs,  
406 and there are concerns about a potentially greater risk of neurotoxicity, our results did not  
407 indicate CRC patients experience a higher rate of SAEs due to having history or evidence upon

408 physical examination of central nervous disease (CNS) disease. If these patients are excluded,  
409 justification for such exclusion should be provided alongside the exclusion criteria (31).

## 410 **Discussion**

411  
412 Rigorous eligibility criteria for RCTs may make the trial participants not representative of real-  
413 world patients. When the trial drugs are on market and prescribed for patients, severe adverse  
414 events could happen. In this paper, we developed a machine learning approach to identify the  
415 patient subgroups from RWD which are more or less likely to encounter SAEs after the  
416 treatment. We consider patient demographics and all clinical events, including diagnosis and  
417 medications, in the baseline period for deriving the subgroups. To account for the high-  
418 dimensionality and SAE information, we proposed a novel supervised topic modeling approach  
419 (SPFA) to achieve this goal. Our approach identified a set of clinical topics and derived novel  
420 patient representations based on them, such that the patient subgroups with or without SAEs  
421 can be well separated with these representations.

422 We validated our method on two RCTs from different disease domains: (1) NCT00478205 for  
423 AD; and (2) NCT00112918 for colorectal cancer. Tested on both trials, patient subgroup  
424 (#SAE=0) and patient subgroup (#SAE>0) can be separated well by k-means clustering using  
425 the inferred topics. The inferred topics characterized as likely to align with the patient subgroup  
426 (#SAE>0) revealed meaningful combinations of clinical features and can provide data-driven  
427 recommendations for refining the exclusion criteria of clinical trials.

428 In a recent study from Liu et al. (32) which aims at evaluating ECs for oncology trials using  
429 RWD and AI, the authors quantified the representability of each study trait with SHAP (33), and  
430 they tried to relax the range of each EC for broadening the participation. Only traits with  
431 continuous values are considered in a one-by-one manner. Our proposed approach in this  
432 paper mainly considered binary traits (continuous traits can also be incorporated with  
433 appropriate discretizations followed by one-hot representations) and modeled the high-order  
434 interactions of these traits as clinical topics. In addition, we also considered adding extra traits to  
435 improve the representability and safety of the trial in real-world data.

436 Our study has several limitations. First, our study only leveraged the RWD from OneFlorida,  
437 which is a regional clinical research network. Future investigation on larger and more diverse  
438 RWD is needed to enhance the generalizability of the identified subgroups. Second, we only  
439 explored structured information in RWD in this study. Lots of important information, such as  
440 symptoms, clinical assessments (e.g., from radiology and pathology reports) and socioeconomic  
441 status, are encoded in clinical notes. Extracting and incorporating unstructured information in  
442 our study is another important direction to pursue. Third, only discrete traits have been  
443 considered in this study. Continuous traits, such as lab tests, are also crucial for many RCTs.  
444 Their corresponding computable counterparts in RWD should be explored as well.

445

446

## 447 **Methods**

448

449 **Clinical trials:** We selected two phase III clinical trials from ClinicalTrials.gov — a registry  
450 maintained by the National Library of Medicine (NLM) in the United States.

451 NCT00478205 is a double-blind, double-dummy trial aims to compare different dosages of  
452 donepezil for treating patients with moderate to severe AD (11). We identified AD patients from  
453 OneFlorida using ICD-9/10-CM codes (i.e., ICD-9: 331.0; ICD-10: G30.0, G30.1, G30.8, and  
454 G30.9). Then, we used the RxNorm concept unique identifier (RXCUI) and National Drug Code  
455 (NDC) codes to identify AD patients who took donepezil.

456 NCT00112918 is a randomized trial studying different combination chemotherapies (i.e. with or  
457 without bevacizumab) in treating CRC patients. Our study focused on the control arm (i.e.,  
458 patients who were treated with the SoC, FOLFOX4). We identified CRC patients using ICD-  
459 9/10-CM codes (i.e., ICD-9: 153.\*, 154.\*, 159.0; ICD-10: C18.\*, C19.\*, C20.\*, C26.0). Within the  
460 CRC patients, we then used the RXCUI and NDC codes of three drugs (i.e., 5-fluorouracil,  
461 leucovorin, and oxaliplatin) (Supplementary Table 3) to further identify the CRC patients who  
462 took FOLFOX4 after their CRC diagnosis.

463 Real-world patient data (RWD) from OneFlorida: We obtained individual-level patient data from  
464 the OneFlorida Clinical Research Consortium (10), which contains robust longitudinal and linked  
465 patient-level RWD of ~15 million (>60%) Floridians, including data from Medicaid claims, cancer  
466 registries, vital statistics, and EHRs from its clinical partners. As one of the Clinical Data  
467 Research Networks (CDRNs) contributing to the national Patient-Centered Clinical Research  
468 Network (PCORnet), OneFlorida includes 12 healthcare organizations that provide care through  
469 4,100 physicians, 914 clinical practices, and 22 hospitals, covering all 67 Florida counties.  
470 OneFlorida follows the PCORnet Common Data Model (CDM) including patient demographics,  
471 enrollment status, vital signs, conditions, encounters, diagnoses, procedures, prescribing (i.e.,  
472 provider orders for medications), dispensing (i.e., outpatient pharmacy dispensing), and lab  
473 results (8).

474 We extracted the clinical events from OneFlorida data, which include demographics (i.e., age,  
475 sex, race), diagnoses (i.e., ICD 9/10), and drugs (i.e., NDC/RXNorm). We calculated age based  
476 on the birth date and first donepezil treatment date for AD patients, or first FOLFOX4 treatment  
477 date for CRC patients. Uniform-sized bins were used to discretize the age and one-hot encoding  
478 was adopted to encode age, gender and race variables. We mapped diagnosis codes to  
479 Phecode which is designed to facilitate phenome-wide association studies (PheWAS) in EHRs  
480 (34). Drug codes (i.e., NDC/RXNorm) were mapped to the Anatomical Therapeutic Chemical  
481 (ATC) Classification System 3rd level. Finally, we concatenated all the features (demographics,  
482 diagnosis and medications) to represent each patient as a binary vector.

483

484 **Defining serious adverse events (SAEs).** To define an SAE, we followed the Food and Drug  
485 Administration (FDA) Code of Federal Regulations (CFR) Title 21 definition of SAE (35). An  
486 adverse event (AE) is considered serious if, in the view of either the investigator or sponsor, it  
487 results in any of the following outcomes: 1) death; 2) a life-threatening adverse event, 3)  
488 inpatient hospitalization or prolongation of existing hospitalization, 4) disability or permanent



489 damage, 5) congenital anomaly/birth defect and 6) important medical events (IME) that may not  
 490 result in death, be life threatening, or require hospitalization may be considered a serious  
 491 adverse drug experience when, based upon medical judgment, they may jeopardize the patient  
 492 or subject and may require medical or surgical intervention to prevent one of the outcomes  
 493 listed in this definition. Another resource for defining SAE is the Common Terminology Criteria  
 494 for Adverse Events (CTCAE) - a descriptive terminology for AE reporting (36). CTCAE  
 495 categorizes AE into Grade 1 (mild), Grade 2 (moderate), Grade 3 (severe or medically  
 496 significant but not immediately life-threatening), Grade 4 (life-threatening consequences), and  
 497 Grade 5 (death). Considering both the definition from FDA CFR title 21 and CTCAE, we defined  
 498 an AE as SAE if it results in hospitalization or death.

499 In this study, to identify SAEs for patients treated with Donepezil or FOLFOX4, we first identified  
 500 the reported SAEs in the Result section of the selected trials from ClinicalTrials.gov. For each  
 501 SAE, we first collected the ICD-9/10-CM codes to identify corresponding health conditions  
 502 (Supplementary Table). To further identify SAEs, we mapped these terms to the CTCAE terms  
 503 and categorized them as SAEs based on the grading scale above (i.e., CTCAE Grade 3/4 and  
 504 5).

505 **Poisson factor analysis.** By collecting all patient vectors, we can construct a binary data matrix  
 506  $X \in \{0, 1\}^{V \times N}$ , with  $V$  corresponding to the number of features and  $N$  being the number of  
 507 patients. Poisson factor analysis (PFA) (37) assumes  $X$  following a Poisson likelihood as

$$X \sim \text{Poisson}(\Phi\theta)$$

508 where  $\Phi = [\phi_1, \dots, \phi_K] \in \mathcal{R}_+^{V \times K}$  is the topic matrix with each column  $\phi_k$  being the  $k$ -th clinical  
 509 topic, and  $\phi_k$  is a distribution over features;  $\theta = [\theta_1, \dots, \theta_N] \in \mathcal{R}_+^{K \times N}$  is the topic weight matrix  
 510 and each column  $\theta_n$  represents the topic weights of the  $n$ -th patient. Based on the expectation  
 511 rule, we have the equation:

$$x_n = \Phi\theta_n = \phi_1\theta_{n1} + \phi_2\theta_{n2} + \dots + \phi_K\theta_{nK}.$$

512 Clearly, each patient vector is composed of weighted summation of all topics, where values in  
 513  $\theta_n$  denotes the weights. Therefore, we call  $\theta_n$  as topic weights, a new representation for  $x_n$ ,  
 514 since it exhibits the weight (or proportion after normalization) of each topic in representing the  
 515 patient  $x_n$ . We will perform clustering on the learned new representations.

516 Compared with latent Dirichlet allocation (LDA) (38), which models the distribution of topic  
 517 weights as Dirichlet distribution, PFA models it as Gamma distribution. The advantage of  
 518 Gamma distribution for topic weight is that it introduces the shrinkage mechanism to prune  
 519 inactive factors and enhances the model interpretability (39). From Fig. 4b and Fig. 6b we can  
 520 see that although we set the number of topics as 40 for both cases, after learning, our model  
 521 automatically truncates it to 35 for AD and 26 for CRC. It is in accordance with the fact that CRC  
 522 has less samples which thus can be described with less topics.  
 523  
 524  
 525

526 **Clustering with supervised PFA models.** The original PFA is purely unsupervised. In order to  
 527 incorporate the outcome information (i.e., having SAE or no) into the topic learning process, we  
 528 extended the original PFA model to a supervised setting which uses the occurrence of SAE as  
 529 the supervision information to guide the learning process of PFA. Specifically, for the  $n$ -th  
 530 patient, if he/she did not encounter any SAE in the follow-up period, we set  $y_n = 0$ ; otherwise,  
 531 we set  $y_n = 1$ . Then we adopted the mean-field variational Bayes method (40) to maximize the  
 532 evidence lower bound (ELBO) of the data likelihood as

$$533 \quad ELBO = \sum_{n=1}^N \mathbb{E}_{q(\theta_n)} [\log p(x_n | \Phi, \theta_n)] - \sum_{n=1}^N \mathbb{E} \left[ \frac{q(\theta_n)}{p(\theta_n | r)} \right],$$

534 where  $p(x_n | \Phi, \theta_n)$  and  $p(\theta_n | r)$  are the Poisson likelihood and Gamma prior as in PFA,  $q(\theta_n)$  is  
 535 the variational posterior to be learned. Currently, we built  $q(\theta_n)$  as an encoder network  
 536  $q_W(\theta_n | x_n)$ , where  $W$  represents learnable parameters of the encoder network, and  $q(\cdot)$  is  
 537 modeled as a Weibull distribution that makes  $\theta_n$  positive and sparse (41).

538 To perform supervised learning, we added a supervised regularizer in the original ELBO  
 539 objective as

$$540 \quad L = ELBO + \sum_{n=1}^N \log p_{W_c}(y_n | \theta_n),$$

541 where the second term can be viewed as the label likelihood implemented by cross entropy loss.  
 542 The model structure of supervised PFA is shown in Figure 1. As a result, we built a probabilistic  
 543 auto-encoding supervised topic model, whose parameters were encoder parameters  $W$ ,  
 544 decoder parameters  $\Phi$  (topics), and classifier  $W_c$ . We deployed stochastic gradient descent to  
 545 learn  $W$  and  $W_c$ , and stochastic gradient based Monte Carlo Markov Chain sampling to infer  $\Phi$   
 546 (41). Our proposed model can be learned in a mini-batch style, which is easily amenable for  
 547 large-scale data analysis.

548 **Mean topic weight (MTW) and Mann–Whitney U (MWU) test.** In our analysis, we use MTW to  
 549 select typical topics and plot the topic distributions on each cluster (pie plot in Fig. 3b).  
 550 According to the data generation process of PFA and equation (1), topic weight of  $n$ -th patient  
 551  $\theta_n$  represents the weights of all topics in representing one patient. For fair evaluation, we  
 552 normalize  $\theta_n$  as  $\tilde{\theta}_n = \theta_n / \sum_k \theta_{nk}$  to a Dirichlet distribution (38). As a result,  $\tilde{\theta}_n$  can be regarded  
 553 as topic proportions. Given a group with  $N'$  patients, the MTW of  $k$ -th topic within this group is  
 554 calculated as  $\sum_{n=1}^{N'} \tilde{\theta}_{nk} / N'$ . For each topic, after calculation of MTW on SAE subgroup and  
 555 non-SAE subgroup, we use MWU test (18) to calculate the p-value of each topic for evaluating  
 556 the significant difference of topic weights on two subgroups.

## 557 **Data availability**

558  
 559 The real data analyzed in this article were provided by OneFlorida Clinical Research  
 560 Consortium and restrictions apply to the availability of these data. Requests for access to the  
 561 data should be submitted to and approved by OneFlorida Clinical Research Consortium.

## 562 **Code availability**

563

564 The implementation of the proposed algorithm in python are publicly available from  
565 <https://github.com/haozhangWCM/Submission-to-NC-SPFA-for-EC>.

## 566 IRB Information

567  
568 This study has been reviewed by the University of Florida Institute Review Board under  
569 protocols IRB202003137 and IRB202000704.

## 570 References

- 571
- 572 1. Hariton E, Locascio JJ. Randomised controlled trials—the gold standard for effectiveness  
573 research. *BJOG: an international journal of obstetrics and gynaecology*. 2018;125(13):1716.
  - 574 2. Kim ES, Bernstein D, Hilsenbeck SG, Chung CH, Dicker AP, Ersek JL, et al. Modernizing  
575 eligibility criteria for molecularly driven trials. *Journal of Clinical Oncology*. 2015;33(25):2815-20.
  - 576 3. Sardar MR, Badri M, Prince CT, Seltzer J, Kowey PR. Underrepresentation of women, elderly  
577 patients, and racial minorities in the randomized trials used for cardiovascular guidelines. *JAMA*  
578 *internal medicine*. 2014;174(11):1868-70.
  - 579 4. Banzi R, Camaioni P, Tettamanti M, Lucca U. Older patients are still under-represented in  
580 clinical trials of Alzheimer's disease. *Alzheimer's research & therapy*. 2016;8(1):1-10.
  - 581 5. Pretorius RW, Gataric G, Swedlund SK, Miller JR. Reducing the risk of adverse drug events in  
582 older adults. *Am Fam Physician*. 2013;87(5):331-6.
  - 583 6. Smits M, Zegers M, Groenewegen PP, Timmermans DR, Zwaan L, van der Wal G, et al.  
584 Exploring the causes of adverse events in hospitals and potential prevention strategies. *Qual Saf*  
585 *Health Care*. 2010;19(5):e5.
  - 586 7. Ory M, Resnick B, Jordan PJ, Coday M, Riebe D, Ewing Garber C, et al. Screening, safety, and  
587 adverse events in physical activity interventions: collaborative experiences from the behavior  
588 change consortium. *Ann Behav Med*. 2005;29 Suppl:20-8.
  - 589 8. Li Q, He Z, Guo Y, Zhang H, George TJ, Hogan W, et al. Assessing the Validity of a a priori  
590 Patient-Trial Generalizability Score using Real-world Data from a Large Clinical Data Research  
591 Network: A Colorectal Cancer Clinical Trial Case Study. *AMIA Annu Symp Proc*.  
592 2019;2019:1101-10.
  - 593 9. FDA-2019-D-1264. Enhancing the Diversity of Clinical Trial Populations — Eligibility Criteria,  
594 Enrollment Practices, and Trial Designs Guidance for Industry. 2019.
  - 595 10. Shenkman E, Hurt M, Hogan W, Carrasquillo O, Smith S, Brickman A, et al. OneFlorida Clinical  
596 Research Consortium: linking a clinical and translational science institute with a community-  
597 based distributive medical education model. *Academic Medicine*. 2018;93(3):451.
  - 598 11. NCT00478205 CgI. Comparison of 23 mg Donepezil Sustained Release (SR) to 10 mg  
599 Donepezil Immediate Release (IR) in Patients With Moderate to Severe Alzheimer's Disease:  
600 <https://clinicaltrials.gov/ct2/show/NCT00478205>; 2007 [
  - 601 12. Institute NC. Common Terminology Criteria for Adverse Events (CTCAE) 2020 [Available from:  
602 [https://ctep.cancer.gov/protocoldevelopment/electronic\\_applications/ctc.htm](https://ctep.cancer.gov/protocoldevelopment/electronic_applications/ctc.htm).
  - 603 13. McInnes L, Healy J, Melville J. UMAP: uniform manifold approximation and projection for  
604 dimension reduction. *arXiv*. arXiv preprint arXiv:1802.03426. 2018.
  - 605 14. Blei DM. Probabilistic topic models. *Communications of the ACM*. 2012;55(4):77-84.
  - 606 15. Syed S, Spruit M, editors. Full-text or abstract? examining topic coherence scores using latent  
607 dirichlet allocation. 2017 IEEE International conference on data science and advanced analytics  
608 (DSAA); 2017: IEEE.

- 609 16. Lleti R, Ortiz MC, Sarabia LA, Sánchez MS. Selecting variables for k-means cluster analysis by  
610 using a genetic algorithm that optimises the silhouettes. *Analytica Chimica Acta*. 2004;515(1):87-  
611 100.
- 612 17. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis.  
613 *Journal of computational and applied mathematics*. 1987;20:53-65.
- 614 18. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger  
615 than the other. *The annals of mathematical statistics*. 1947:50-60.
- 616 19. Li Y, Nair P, Lu XH, Wen Z, Wang Y, Dehaghi AAK, et al. Inferring multimodal latent topics from  
617 electronic health records. *Nature communications*. 2020;11(1):1-17.
- 618 20. Gerretsen P, Pollock BG. Drugs with anticholinergic properties: a current perspective on use and  
619 safety. *Expert opinion on drug safety*. 2011;10(5):751-65.
- 620 21. Long J, Zhang C-J, Zhu N, Du K, Yin Y-F, Tan X, et al. Lipid metabolism and carcinogenesis,  
621 cancer development. *American journal of cancer research*. 2018;8(5):778.
- 622 22. Di Stadio A, Ralli M, Roccamatysi D, Scarpa A, Della Volpe A, Cassandro C, et al. Hearing loss  
623 and dementia: radiologic and biomolecular basis of their shared characteristics. A systematic  
624 review. *Neurological Sciences*. 2021:1-10.
- 625 23. Wiels W, Baeken C, Engelborghs S. Depressive symptoms in the elderly—An early symptom of  
626 dementia? A systematic review. *Frontiers in pharmacology*. 2020;11:34.
- 627 24. Thirunavukkarasu S, Khader SA. Advances in cardiovascular disease lipid research can provide  
628 novel insights into mycobacterial pathogenesis. *Frontiers in cellular and infection microbiology*.  
629 2019;9:116.
- 630 25. Dhondup T, Qian Q. Acid-base and electrolyte disorders in patients with and without chronic  
631 kidney disease: an update. *Kidney Diseases*. 2017;3(4):136-48.
- 632 26. Samper-Ternent R, Al Snih S. Obesity in older adults: epidemiology and implications for disability  
633 and disease. *Reviews in clinical gerontology*. 2012;22(1):10-34.
- 634 27. Chait MM. Gastroesophageal reflux disease: Important considerations for the older patients.  
635 *World journal of gastrointestinal endoscopy*. 2010;2(12):388.
- 636 28. Baldassarre JS, Kaye D. Special problems of urinary tract infection in the elderly. *Medical Clinics*  
637 *of North America*. 1991;75(2):375-90.
- 638 29. Rubin DC, Shaker A, Levin MS. Chronic intestinal inflammation: inflammatory bowel disease and  
639 colitis-associated colon cancer. *Frontiers in immunology*. 2012;3:107.
- 640 30. Katz S, Feldstein R. Inflammatory bowel disease of the elderly: a wake-up call. *Gastroenterology*  
641 *& hepatology*. 2008;4(5):337.
- 642 31. Kim ES, Bruinooge SS, Roberts S, Ison G, Lin NU, Gore L, et al. Broadening Eligibility Criteria to  
643 Make Clinical Trials More Representative: American Society of Clinical Oncology and Friends of  
644 Cancer Research Joint Research Statement. *J Clin Oncol*. 2017;35(33):3737-44.
- 645 32. Liu R, Rizzo S, Whipple S, Pal N, Pineda AL, Lu M, Arnieri B, Lu Y, Capra W, Copping R, Zou J.  
646 Evaluating eligibility criteria of oncology trials using real-world data and AI. *Nature*. 2021 Apr 7:1-  
647 5.
- 648 33. Lundberg S, Lee SI. A unified approach to interpreting model predictions. arXiv preprint  
649 arXiv:1705.07874. 2017 May 22.
- 650 34. Wu P, Gifford A, Meng X, Li X, Campbell H, Varley T, et al. Mapping ICD-10 and ICD-10-CM  
651 Codes to Phecodes: Workflow Development and Initial Evaluation. *JMIR Med Inform*.  
652 2019;7(4):e14325.
- 653 35. U.S. Food & Drug Administration (FDA). CFA - Code of Federal Regulations: Title 21: Food and  
654 drugs [updated Apr 1, 2020]. Available from:  
655 <https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/cfrsearch.cfm?fr=312.32>.

- 656 36. National Cancer Institute. Common Terminology Criteria for Adverse Events (CTCAE) 2020  
657 [updated Sep 21, 2020]. Available from:  
658 [https://ctep.cancer.gov/protocoldevelopment/electronic\\_applications/ctc.htm](https://ctep.cancer.gov/protocoldevelopment/electronic_applications/ctc.htm).  
659 37. Zhou M, Hannah L, Dunson D, Carin L, editors. Beta-negative binomial process and Poisson  
660 factor analysis. Artificial Intelligence and Statistics; 2012: PMLR.  
661 38. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. the Journal of machine Learning research.  
662 2003;3:993-1022.  
663 39. Zhang H, Chen B, Cong Y, Guo D, Liu H, Zhou M. Deep autoencoding topic model with scalable  
664 hybrid Bayesian inference. IEEE Transactions on Pattern Analysis and Machine Intelligence.  
665 2020.  
666 40. Jordan MI, Ghahramani Z, Jaakkola TS, Saul LK. An introduction to variational methods for  
667 graphical models. Machine learning. 1999;37(2):183-233.  
668 41. Zhang H, Chen B, Guo D, Zhou M. WHAI: Weibull hybrid autoencoding inference for deep topic  
669 modeling. arXiv preprint arXiv:180301328. 2018.  
670

## 671 **Acknowledgements**

672  
673 The work from JX, HZ and FW was supported by NSF 1750326, ONR N00014-18-1-2585 and  
674 NIH RF1AG072449. The work from JB and HZ was supported by NIH R21AG068717 and NIH  
675 R21CA253394.

## 676 **Author contributions**

677  
678 FW and BJ conceived the study. JX and HZ designed and implemented the algorithm. HZ  
679 analyzed the data. JX and HZ drafted the manuscript. All authors polished and proofread the  
680 manuscript.

## 681 **Competing interests**

682  
683 None declared.

684  
685  
686  
687  
688  
689  
690  
691  
692

693 **Figure Legends**

694

695 **Fig. 1: Model overview.** Clinical events including demographics, diagnoses, and medications  
696 were extracted from RWD to represent patients. Supervised Poisson factor analysis (PFA) is  
697 applied to identify patient subgroups with coherent clinical latent topics and outcomes measured  
698 by SAEs. Subgroups with less SAEs can be derived to inform the design of clinical trial eligibility  
699 criteria.

700

701 **Fig. 2: The donepezil trial for Alzheimer’s disease.** **a** Definition of key dates; **b** Selection of  
702 target population; **c** Traits distribution with UMAP among two patients’ groups with or without  
703 SAE.

704 **Fig. 3: Clustering results of AD target population.** **a** Data visualization with UMAP; **b** Topic  
705 distributions of each cluster; **c** Top clinical features from certain disease topics, where the red,  
706 blue, and green topics represent the commonly used topics for clusters 4, 5, 6 (most of samples  
707 are with SAE>0), clusters 1, 2, 3 (most of samples are with SAE=0), and all clusters,  
708 respectively.

709

710 **Fig. 4: Donepezil clinical trial.** **a** Visualization of clustering results; **b** Mean topic weight (MTW)  
711 of all topics on two groups, where x-axis is the topic index and y-axis is the MTW of each topic  
712 on two subgroups.

713

714 **Fig. 5: The Bevacizumab (Avastin) clinical trial for colon cancer.** **a** Definition of key dates; **b**  
715 Selection of the target population; **c** Traits distribution with UMAP among the two patients’  
716 groups with or without SAEs.

717

718 **Fig. 6: Bevacizumab (Avastin) clinical trial.** **a** Visualization of clustering results; **b** Mean topic  
719 weight for 40 topics where x-axis is the topic and y-axis is the proportion of each group in that  
720 topic.

721

722 **Fig. 7: Top clinical features from disease topics in Bevacizumab (Avastin) clinical trial.**

723

724

725  
726

**Table 1.** Demographic characteristics and selected traits of the target population of the donepezil trial for Alzheimer's disease.

| <b>Characteristic</b>                        | <b>Overall<br/>(N=4,998)</b> | <b># SAEs = 0<br/>(N=3,063)</b> | <b># SAEs &gt; 0<br/>(N=1,935)</b> | <b><math>\chi^2</math><br/>p-value</b> | <b>Adjust Demo*</b> |
|--|------------------------------|---------------------------------|------------------------------------|--|---------------------|
| <b>Age, Mean (SD), yr</b>                    | 77.53 (9.9)                  | 76.98 (9.8)                     | 78.41 (9.9)                        |  |                     |
| <b>Sex, No. (%)</b>                          |                              |                                 |                                    |  |                     |
| <b>Female</b>                                | 3,123 (62.5)                 | 1,923 (62.8)                    | 1,200 (62.0)                       |  |                     |
| <b>Race/Ethnicity, No. (%)</b>               |                              |                                 |                                    |  |                     |
| White  | 3,537 (70.8)                 | 2,262 (73.8)                    | 1,275 (65.8)                       |  |                     |
| Black  | 965 (19.4)                   | 484 (158)                       | 481 (24.9)                         |  |                     |
| Asian  | 34 (0.6)                     | 25 (0.8)                        | 9 (0.5)                            |  |                     |
| Others & Unknown                             | 462 (9.2)                    | 292 (9.5)                       | 170 (8.8)                          |  |                     |
| <b>Study traits, No. (%)</b>                 |                              |                                 |                                    |  |                     |
| Memantine                                    | 1,511 (30.2)                 | 950 (31)                        | 561 (28.9)                         | 0.129                                  | 0.145               |
| Psychiatric disorders                        | 1,396 (27.9)                 | 754 (24.6)                      | 642 (33.1)                         | ≤0.001                                 | ≤0.001              |
| Cardiovascular (CS*)                         | 1,082 (21.6)                 | 492 (16)                        | 590 (30.4)                         | ≤0.001                                 | ≤0.001              |
| Endocrine (CS*)                              | 813 (16.2)                   | 353 (11.5)                      | 460 (23.7)                         | ≤0.001                                 | ≤0.001              |
| Cancer                                       | 808 (16.1)                   | 468 (15.2)                      | 340 (17.5)                         | 0.032                                  | 0.091               |
| Dysphagia                                    | 649 (12.9)                   | 302 (9.8)                       | 347 (17.9)                         | ≤0.001                                 | ≤0.001              |
| Gastrointestinal (CS*)                       | 631 (12.6)                   | 289 (9.4)                       | 342 (17.6)                         | ≤0.001                                 | ≤0.001              |
| Drug or alcohol abuse<br>or dependence       | 627 (12.5)                   | 283 (9.2)                       | 344 (17.7)                         | ≤0.001                                 | ≤0.001              |
| Respiratory (CS*)                            | 586 (11.7)                   | 249 (8.1)                       | 337 (17.4)                         | ≤0.001                                 | ≤0.001              |
| AD with delirium                             | 389 (7.7)                    | 161 (5.2)                       | 228 (11.7)                         | ≤0.001                                 | ≤0.001              |
| Hepatic disease                              | 361 (7.2)                    | 180 (5.8)                       | 181 (9.3)                          | ≤0.001                                 | ≤0.001              |
| Renal (CS*)                                  | 342 (6.8)                    | 135 (4.4)                       | 207 (10.6)                         | ≤0.001                                 | ≤0.001              |
| Parkinson disease                            | 329 (6.5)                    | 176 (5.7)                       | 153 (7.9)                          | 0.003                                  | 0.001               |
| Menopausal                                   | 230 (4.6)                    | 128 (4.1)                       | 102 (5.2)                          | 0.073                                  | 0.040               |
| Antidepressant                               | 226 (4.5)                    | 143 (4.6)                       | 83 (4.2)                           | 0.530                                  | 0.590               |
| Basal/squamous<br>cell carcinoma of the skin | 216 (4.3)                    | 128 (4.1)                       | 88 (4.5)                           | 0.532                                  | 0.275               |
| Gastric ulcers                               | 163 (3.2)                    | 75 (2.4)                        | 88 (4.5)                           | ≤0.001                                 | ≤0.001              |
| Inflammatory bowel disease                   | 154 (3)                      | 82 (2.6)                        | 72 (3.7)                           | 0.038                                  | 0.024               |
| Rivastigmine                                 | 153 (3)                      | 121 (3.9)                       | 32 (1.6)                           | ≤0.001                                 | ≤0.001              |
| Multi-infarct dementia                       | 151 (3)                      | 72 (2.3)                        | 79 (4)                             | ≤0.001                                 | 0.001               |
| Acupressure                                  | 119 (2.3)                    | 54 (1.7)                        | 65 (3.3)                           | ≤0.001                                 | 0.002               |
| Fecal incontinence                           | 107 (2.1)                    | 48 (1.5)                        | 59 (3)                             | ≤0.001                                 | 0.001               |
| Galantamine                                  | 35 (0.7)                     | 18 (0.5)                        | 17 (0.8)                           | 0.230                                  | 0.190               |
| Severe lactose intolerance                   | 24 (0.4)                     | 12 (0.3)                        | 12 (0.6)                           | 0.255                                  | 0.219               |
| Hepatic (CS*)                                | 20 (0.4)                     | 9 (0.2)                         | 11 (0.5)                           | 0.134                                  | 0.105               |

727  
728

CS\*: Clinically significant. If the disease causes hospitalization, we consider it as "clinically significant". Demo\*: Demographics (i.e., age, gender and race).

729 **Table 2.** Demographic characteristics and selected traits of the target population of the  
 730 bevacizumab (Avastin) clinical trial for colorectal cancer.

| <b>Characteristic</b>             | <b>Overall<br/>(N=739)</b> | <b># SAEs = 0<br/>(N=347)</b> | <b># SAEs &gt; 0<br/>(N=392)</b> | <b><math>\chi^2</math><br/>p-value</b> | <b>Adjust Demo*</b> |
|-----------------------------------|----------------------------|-------------------------------|----------------------------------|--|---------------------|
| <b>Age, Mean (SD), yr</b>         | 57.49 (11.2)               | 59.13 (11.2)                  | 56.0 (11.1)                      |  |                     |
| <b>Sex, No. (%)</b>               |                            |                               |                                  |  |                     |
| Female                            | 328 (44.3)                 | 141 (40.6)                    | 187 (47.7)                       |  |                     |
| <b>Race/Ethnicity, No. (%)</b>    |                            |                               |                                  |  |                     |
| White                             | 488 (66.0)                 | 237 (68.3)                    | 251 (64)                         |  |                     |
| Black                             | 172 (23.3)                 | 79 (22.8)                     | 93 (23.7)                        |  |                     |
| Asian                             | 10 (1.4)                   | 5 (1.4)                       | 5 (1.2)                          |  |                     |
| Others & Unknown                  | 69 (9.3)                   | 26 (7.5)                      | 43 (10.9)                        |  |                     |
| <b>traits, No. (%)</b>            |                            |                               |                                  |  |                     |
| Colon carcinoma                   | 616 (83.3)                 | 296 (85.3)                    | 320 (81.6)                       | 0.182                                  | 0.058               |
| Metastatic disease                | 499 (67.5)                 | 221 (63.6)                    | 278 (70.9)                       | 0.036                                  | 0.026               |
| parenteral anticoagulants         | 240 (32.4)                 | 77 (22.1)                     | 163 (41.5)                       | $\leq 0.001$                           | $\leq 0.001$        |
| immunotherapy                     | 146 (19.7)                 | 75 (21.6)                     | 71 (18.1)                        | 0.233                                  | 0.130               |
| anti-angiogenic treatment         | 137 (18.5)                 | 72 (20.7)                     | 65 (16.5)                        | 0.146                                  | 0.081               |
| myocardial infarction             | 90 (12.1)                  | 25 (7.2)                      | 65 (16.5)                        | $\leq 0.001$                           | $\leq 0.001$        |
| Significant traumatic injury      | 40 (5.4)                   | 16 (4.6)                      | 24 (6.1)                         | 0.365                                  | 0.379               |
| thrombolytic agent                | 38 (5.1)                   | 9 (2.5)                       | 29 (7.3)                         | 0.003                                  | 0.003               |
| central nervous disease           | 37 (5)                     | 12 (3.4)                      | 25 (6.3)                         | 0.070                                  | 0.099               |
| inability to take oral medication | 33 (4.4)                   | 10 (2.8)                      | 23 (5.8)                         | 0.050                                  | 0.034               |
| Open biopsy                       | 30 (4)                     | 17 (4.8)                      | 13 (3.3)                         | 0.277                                  | 0.282               |
| radiotherapy                      | 24 (3.2)                   | 10 (2.8)                      | 14 (3.5)                         | 0.598                                  | 0.790               |
| bone fracture                     | 21 (2.8)                   | 8 (2.3)                       | 13 (3.3)                         | 0.410                                  | 0.392               |
| coagulopathy                      | 20 (2.7)                   | 8 (2.3)                       | 12 (3)                           | 0.528                                  | 0.672               |
| oophorectomy                      | 17 (2.3)                   | 7 (2)                         | 10 (2.5)                         | 0.630                                  | 0.956               |
| cerebrovascular accidents         | 14 (1.8)                   | 5 (1.4)                       | 9 (2.2)                          | 0.396                                  | 0.428               |

731

732