**数据结构：**

我的数据按年、季度分类，每个季度的数据集包括demo，drug，indi，outc，reac，rpsr及ther7个文件。每个季度的表格相同，字段相同，唯一区别是具体数据发生了变化。

其中demo表代表病人基本信息，primaryid和caseid均是唯一编号，该表有价值的字段包括event\_dt，age，sex，wt，occp\_cod，reporter\_country，代表事件发生时间，年龄，性别，体重，报告人职业，报告国家。

drug表表示使用药物信息，primaryid和caseid不是唯一值，如果一个患者同时使用了多种药物，药物将按照drug\_seq按顺序编号，同时对其中药物是否是导致不良反应的元凶进行定性判断，该表有价值的字段包括drug\_seq，role\_cod，drugname，route，dechal，rechal，dose\_amt，dose\_unit，dose\_form，dose\_freq，代表药物序号，角色，药名，给药途径，给药剂量，给药形式，给药频次。

indi表表示药物适应症信息，primaryid和caseid不是唯一值，如果一个患者同时使用了多种药物，药物将按照indi\_drug\_seq按顺序编号，分别对应其治疗适应症，该表有用字段包括indi\_drug\_seq，代表该药在治疗过程中用来做什么的。

outc表表示治疗结局，primaryid和caseid不是唯一值，结局有多个值，可能代表不同阶段的结局，难以判断结局时间谁先谁后，该表有用字段包括outc\_cod，表示治疗结局，包括DE，LT，HO，DS，CA，RT，OT7个值，分表表示死亡，危及生命，需要住院治疗，残疾，认知功能障碍，需要干预及其他。

reac表表示发生的不良反应，primaryid和caseid不是唯一值，因为可能出现多种不良反应，但难以判断这些不良反应与哪个药物对应，也难以判断这些不良反应导致了哪个结局。有用的字段为pt。

rpsr代表报告来源。

ther表表示治疗信息。primaryid和caseid不是唯一值，因为患者使用多种药物，每种药物的疗程均与dsg\_drug\_seq对应。该表有用的字段包括dur和dur\_cod，表示疗程。

7个文件中有3个主要字段可用于连接各个表，如下

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **demo** | **outc** | **reac** | **rpsr** | **drug** | **indi** | **Ther** |
| primaryid | primaryid | primaryid | primaryid | primaryid | primaryid | primaryid |
| caseid | caseid | caseid | caseid | caseid | caseid | caseid |
|  |  |  |  | drug\_seq | indi\_drug\_seq | dsg\_drug\_seq |

*\*其中drug\_seq，indi\_drug\_seq，dsg\_drug\_seq代表同一个意思。*

**需要提取的字段包括但不限于：primaryid，caseid，drug\_seq，role\_cod，pt，rechal，dechal，lit\_ref，indi\_pt，i\_f\_code，age，age\_cod，age\_grp，sex，wt，wt\_cod，drugname，dose\_form，dose\_freq，dose\_unit，cum\_dose\_chr，cum\_dose\_unit，route，start\_dt，start\_dt\_num，end\_dt，end\_dt\_num，dur，dur\_cod，occp\_cod，occr\_country，outc\_cod，event\_dt，reporter\_country，rpsr\_cod（如有可能全部提取最佳）。**

**我希望得到的功能包括：**

**功能一：**

1. 检索drug2018q1.csv表中drugname包含Risperidone（不区分大小写），且role\_cod等于PS或SS的记录，筛选出该表中的primaryid，caseid，drug\_seq，role\_cod，drugname，prod\_ai，val\_vbm，route，dose\_vbm，cum\_dose\_chr，cum\_dose\_unit，dechal，rechal，lot\_num，exp\_dt，exp\_dtstr，exp\_dt\_mult，nda\_num，dose\_amt，dose\_unit，dose\_form，dose\_freq字段汇总成表drug1。
2. 以1中筛选到得到的caseid和drug\_seq（该字段等于indi表中的indi\_drug\_seq）检索indi2018q1.csv表中的数据，提取出该表中的primaryid，caseid，indi\_drug\_seq，indi\_pt字段会中成表indi1。
3. 以1中筛选到得到的caseid和drug\_seq（该字段等于ther表中的dsg\_drug\_seq）检索ther2018q1.csv表中的数据，提取出该表中的primaryid，caseid，dsg\_drug\_seq，start\_dt，start\_dt\_num，end\_dt，end\_dt\_num，dur，dur\_cod字段汇总成表ther1。
4. 以1中筛选到得到的caseid检索demo2018q1.csv表中的数据，提取该表中的primaryid，caseid，caseversion，i\_f\_code，i\_f\_code\_num，event\_dt，event\_dt\_num，mfr\_dt，mfr\_dt\_num，init\_fda\_dt，init\_fda\_dt\_num，fda\_dt，fda\_dt\_num，rept\_cod，rept\_cod\_num，auth\_num，mfr\_num，mfr\_sndr，lit\_ref，age，age\_cod，age\_grp，age\_grp\_num，sex，e\_sub，wt，wt\_cod，rept\_dt，rept\_dt\_num，to\_mfr，occp\_cod，reporter\_country，occr\_country，occp\_cod\_num字段汇总成表demo1。
5. 以1中筛选到得到的caseid检索outc2018q1.csv表中的数据，提取该表中的primaryid，caseid，outc\_cod字段汇总成表outc1。
6. 以1中筛选到得到的caseid检索reac2018q1.csv表中的数据，提取该表中的primaryid，caseid，pt字段汇总成表reac。
7. 以1中筛选到得到的caseid检索rpsr2018q1.csv表中的数据，提取该表中的primaryid，caseid，rpsr\_cod字段汇总成表rpsr1。

**功能二：**

1. 检索reac2018q1.csv表中pt字段包含Neurotoxicity（不区分大小写）的记录，提取该表中的primaryid，caseid，pt字段汇总成表reac2。
2. 以1中筛选到得到的caseid检索outc2018q1.csv表中的数据，提取该表中的primaryid，caseid，outc\_cod字段汇总成表outc2。
3. 以1中筛选到得到的caseid检索demo2018q1.csv表中的数据，提取该表中的primaryid，caseid，caseversion，i\_f\_code，i\_f\_code\_num，event\_dt，event\_dt\_num，mfr\_dt，mfr\_dt\_num，init\_fda\_dt，init\_fda\_dt\_num，fda\_dt，fda\_dt\_num，rept\_cod，rept\_cod\_num，auth\_num，mfr\_num，mfr\_sndr，lit\_ref，age，age\_cod，age\_grp，age\_grp\_num，sex，e\_sub，wt，wt\_cod，rept\_dt，rept\_dt\_num，to\_mfr，occp\_cod，reporter\_country，occr\_country，occp\_cod\_num字段汇总成表demo2。
4. 以1中筛选到得到的caseid检索rpsr2018q1.csv表中的数据，提取该表中的primaryid，caseid，rpsr\_cod字段汇总成表rpsr2。
5. 以1中筛选到得到的caseid及drug2018q1.csv表中role\_cod等于PS或SS，检索drug2018q1.csv表，筛选出该表中的primaryid，caseid，drug\_seq，role\_cod，drugname，prod\_ai，val\_vbm，route，dose\_vbm，cum\_dose\_chr，cum\_dose\_unit，dechal，rechal，lot\_num，exp\_dt，exp\_dtstr，exp\_dt\_mult，nda\_num，dose\_amt，dose\_unit，dose\_form，dose\_freq字段汇总成表drug2。
6. 以5中筛选到得到的caseid和drug\_seq（该字段等于indi表中的indi\_drug\_seq）检索indi2018q1.csv表中的数据，提取出该表中的primaryid，caseid，indi\_drug\_seq，indi\_pt字段会中成表indi2。
7. 以5中筛选到得到的caseid和drug\_seq（该字段等于ther表中的dsg\_drug\_seq）检索ther2018q1.csv表中的数据，提取出该表中的primaryid，caseid，dsg\_drug\_seq，start\_dt，start\_dt\_num，end\_dt，end\_dt\_num，dur，dur\_cod字段汇总成表ther2。

因为数据量很大，我只给了其中的一个季度的数据，因此我希望能知道处理的软件及详细代码，后期我可以通过简单修改部分代码就可以处理其他季度的数据。另我对R语言懂一点点，如果R语言能处理最佳。

谢谢！

2018年11月5日