Methods：

1. 从sav文件中统计了包括dfstroke（医生诊断结果）和另外86项颈部血管筛查数据中有效数据的个数，显然dfstroke都是有效数据，另外86项中的有效数据个数见文件valid data number for each item.csv。
2. 从第一步的统计结果可以看出，绝大多数的检查项目有效数据都超过了58800，只有16个是低于100的，这里对于其余的86-16+1项中有效数据对应的病人id求交，即查看哪些病人的71项数据都处于有效状态，准备后续分析，结果显示共有58787个病人的数据是完全有效的。结果保存在两个文件中，第一个是data with 70 features.csv，描述了dfstroke以及70个检测项结果；第二个文件是data with 86 features.csv，描述了dfstroke和所有的86个检测项。
3. 这一步分析了每一项的检查结果有哪些可能的数值，以便于在做深度学习时实现有效的映射到【0,1】区间。结果见range of each examination.csv。经分析，共有约2.5\*10^40种可能的组合。
4. 考虑到之前分析时出现了冲突数据的问题，即，检测结果完全一样而医生诊断结果不一样，在这里分析了一下有多少种检测结果。发现58787组数据中只有19581种检测结果，也就是说，平均大约每个人都能找到另外两个人和自己的检测结果是一样的。结果保存在unique data of the examination.csv中。进一步，分析了一下包括医生的诊断结果在内，会有多少种彼此不同的情况出现，发现有21713种，结果保存在unique data of the examination and diagnose.csv中，这也跟我们之前得到的结果一致——确实有些病人的检测结果一下，但是医生诊断不同。

Suggestions：

1. 去掉缺失比较严重的数据，因为我们确实没有这个数据，全都设置成0逻辑上也说得过去，但如果这些数据对模型本该有影响但是缺失了，就不好解释了。
2. 去掉冲突数据，相同检测结果得到不同诊断，会影响学习过程，建议去掉。
3. 合并重复数据，重复数据比例太大，在划分训练集和测试集时不能保证独立性，对学习的评价很大程度上是在和训练集相同的数据上进行的，评价没有意义。