作业四-ProM流程挖掘

522031910213 朱涵

1 数据预处理

选择了第6个数据表。

编写python程序进行数据的预处理，包括取开始时间和结束时间的平均值作为事件时间、过滤掉所有“挂号”事件并非第一个事件的流程（正常的流程显然都是从挂号开始的），最后导出csv文件，导入到ProM中转换为xes文件。

Python程序如图：

```python
# 读取 Excel 文件
df = pd.read_excel("data.xlsx")
# 确保 START 和 END 列是 datetime 类型
df["ACTIVITY_START"] = pd.to_datetime(df["ACTIVITY_START"], format="%Y-%m-%d %H:%M:%S")
df["ACTIVITY_END"] = pd.to_datetime(df["ACTIVITY_END"], format="%Y-%m-%d %H:%M:%S")

# 计算 START 和 END 的平均值作为 TIME 列
df["TIME"] = df.apply(
    lambda row: (
        row["ACTIVITY_START"] + (row["ACTIVITY_END"] - row["ACTIVITY_START"]) / 2
    ).strftime("%Y-%m-%d %H:%M:%S"),
    axis=1,
)
saved_data_list = []
other_data_list = []
# 过滤掉所有“第一个事件不是挂号”的数据
def filter_by_activity(group: pd.DataFrame) -> pd.DataFrame:
    group = group.sort_values(by="ACTIVITY_START")
    first_activity = group.iloc[0]["ACTIVITY"]

    if first_activity != "挂号":
        other_data_list.append(group)
    else:
        saved_data_list.append(group)

    return group
# 处理并保存结果
df.groupby("GUAHAO_ID").apply(filter_by_activity)
# 合并并保存数据
saved_data = pd.concat(saved_data_list, ignore_index=True)
saved_data.to_csv("processed_data.csv", index=False)
if len(other_data_list) > 0:
    other_data = pd.concat(other_data_list, ignore_index=True)
    other_data.to_csv("other_data.csv", index=False)
```

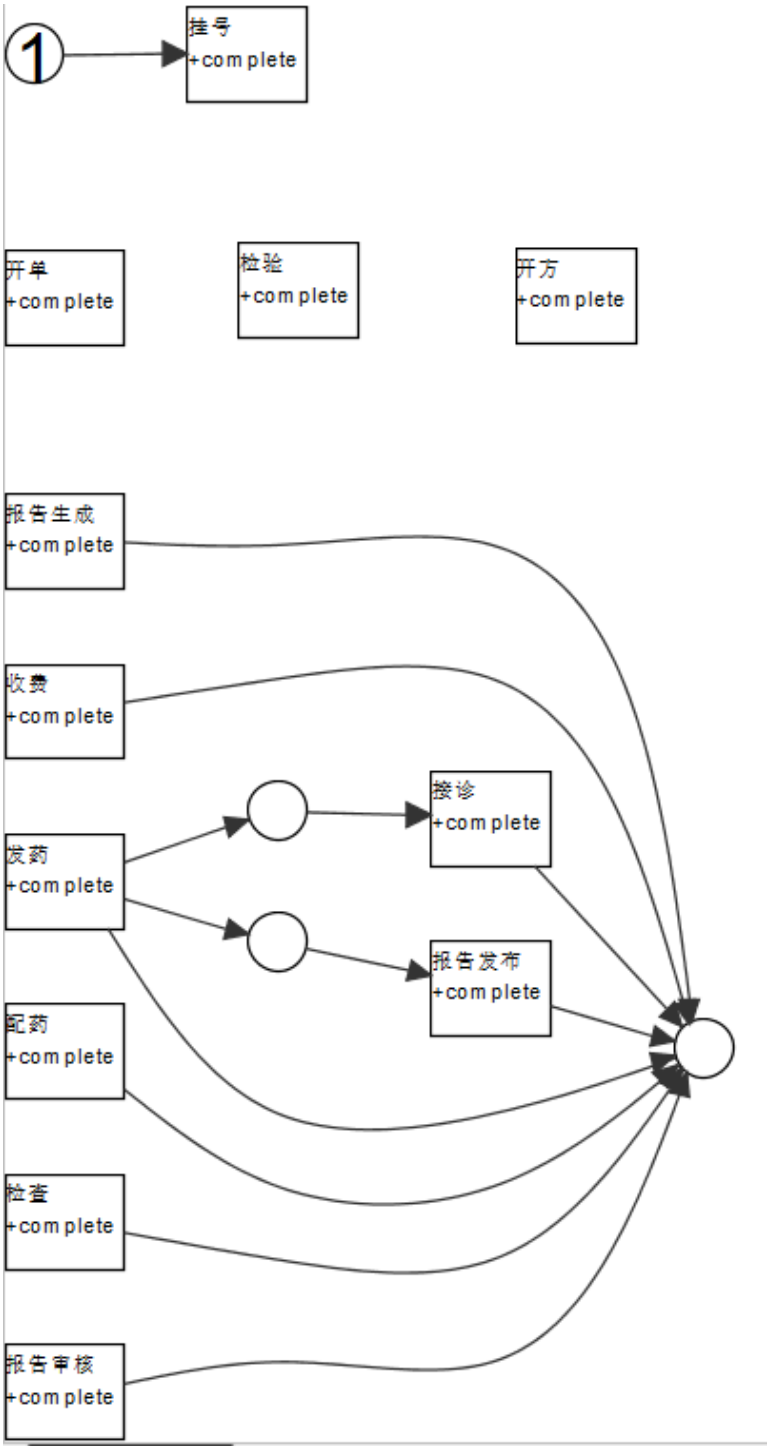
转换为xes文件后，ProM内的统计如下图：

可以看到流程最少有4个事件，最多有34个，其中4个事件的流程最多，有12个流程。

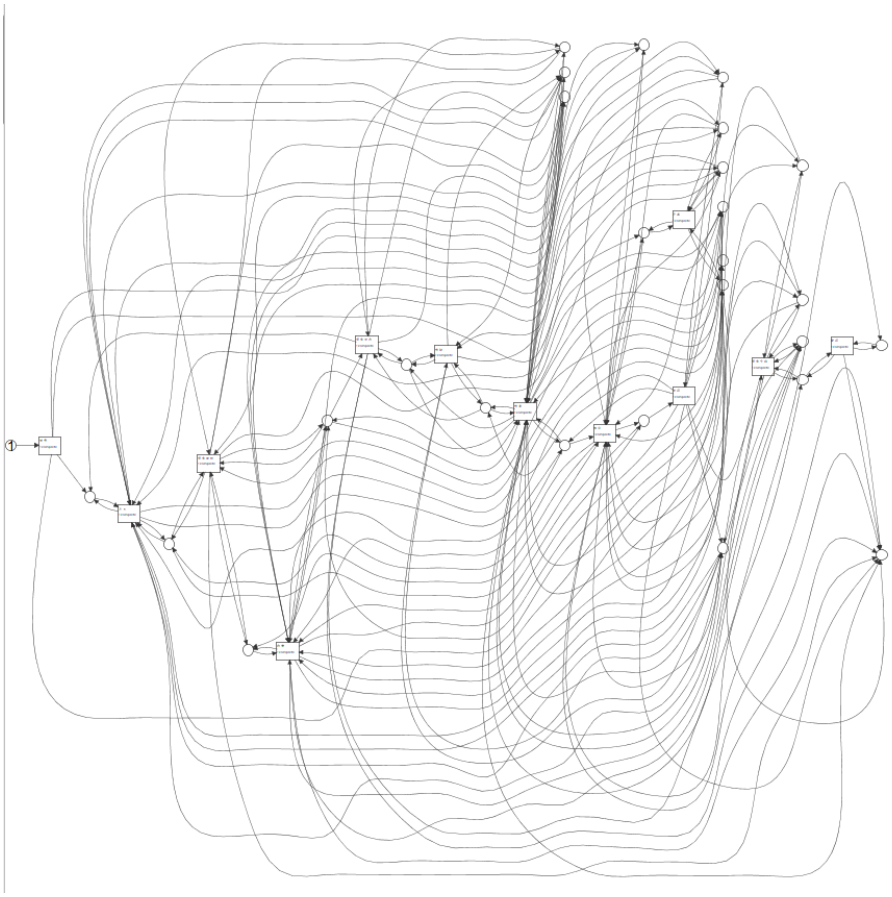
2 流程挖掘

2.1 Alpha算法

Alpha算法是流程挖掘中最经典的一种算法，通过对事件集合的频繁集进行挖掘得出模型，广泛用于从日志中自动推导出流程控制结构。采用Alpha算法挖掘得到模型如下图：



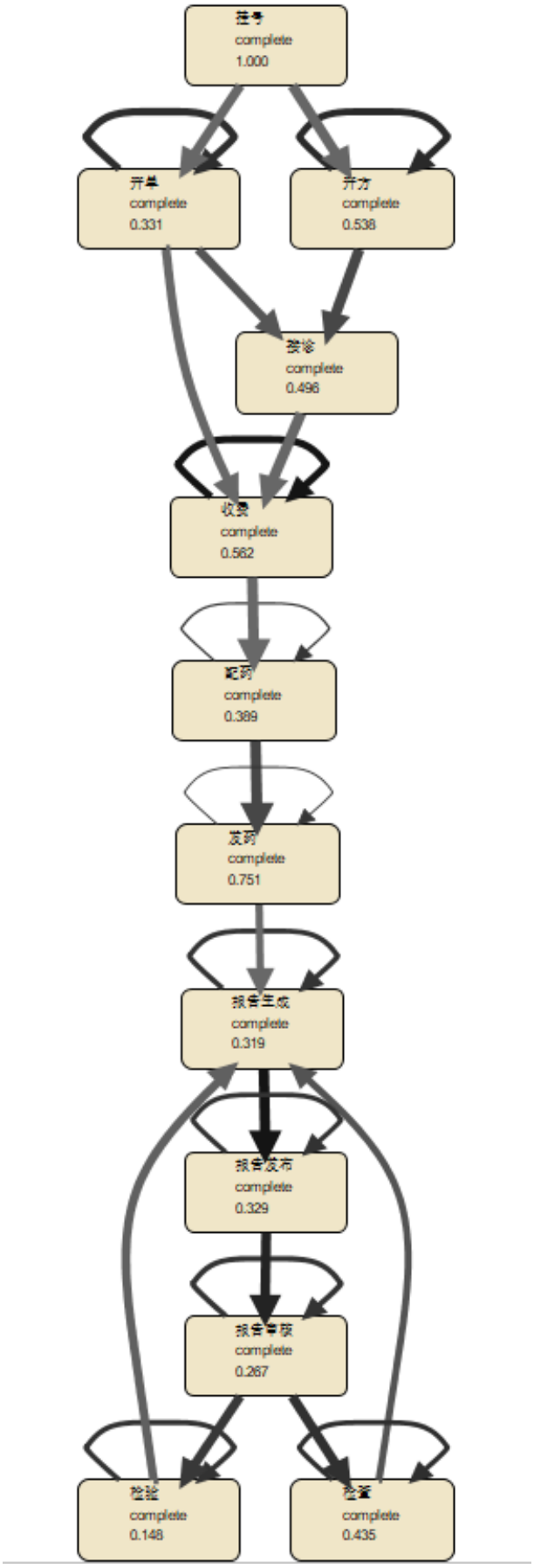
可以看到，得到的模型中还有4个事件是独立的，没有产生连接。采用Alpha+的结果基本类似。下面采用**Alpha++算法**，得到结果如图：



可以看出模型过于复杂，产生了太多连接，推测原因是流程数据中存在太多特例数据（出现次数比较少的数据），产生了过多特例的连接。

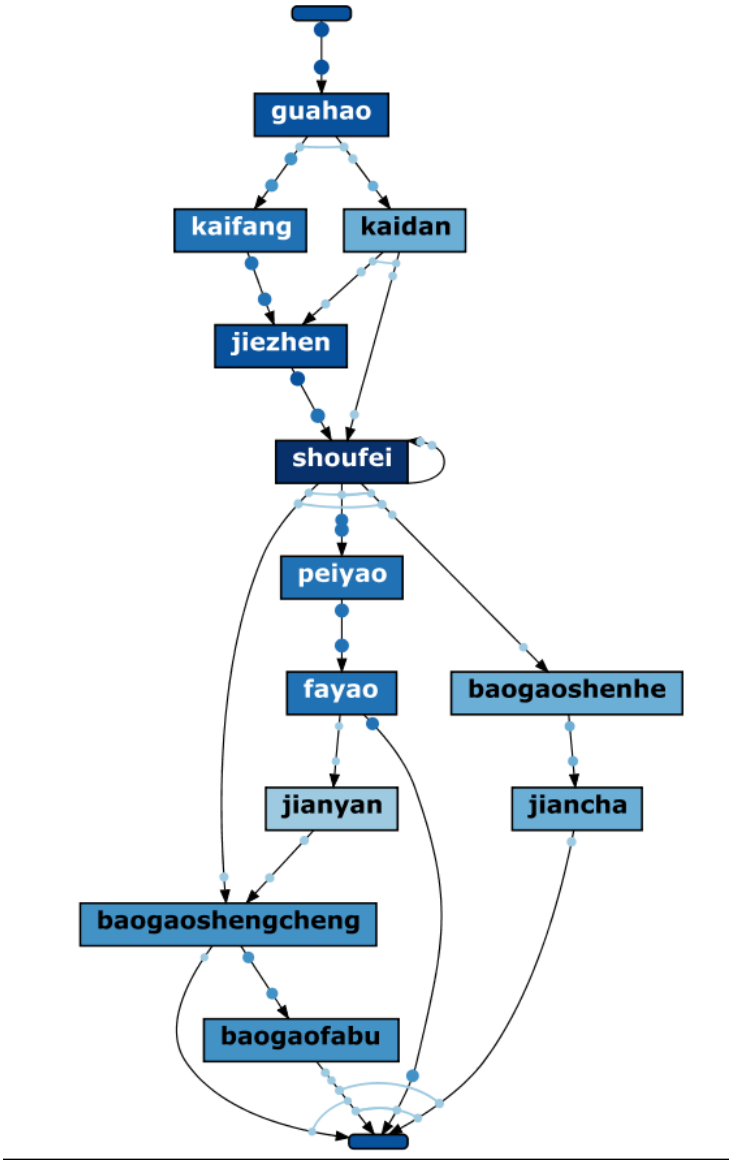
2.2 Fuzzy算法

Fuzzy算法（模糊算法）基于模糊集合理论，通过模糊化的概念，将现实世界中的模糊信息转化为数学模型，从而在决策、控制和分类等领域进行推理和判断。与传统的精确算法不同，Fuzzy算法允许输入和输出在一定范围内具有模糊性，而非绝对的“是”或“否”，通过模糊规则进行推理，最终得到近似的结果。下面是Fuzzy算法得到的模型：

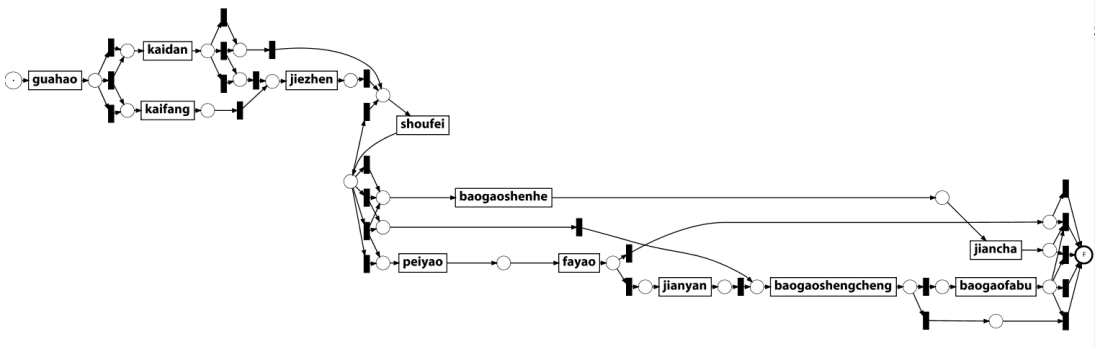

挂号
complete
1.000
开单
complete
0.331
开方
complete
0.538
就诊
complete
0.496
收费
complete
0.562
配药
complete
0.389
发药
complete
0.751
报告生成
complete
0.319
报告发布
complete
0.329
报告审核
complete
0.267
检验
complete
0.148
检查
complete
0.435

2.3 启发式算法

启发式算法（Heuristic Miner）利用启发式规则和局部搜索策略分析事件之间的关系和行为模式，自动生成一个描述实际流程的模型。下面是启发式算法得到的模型：

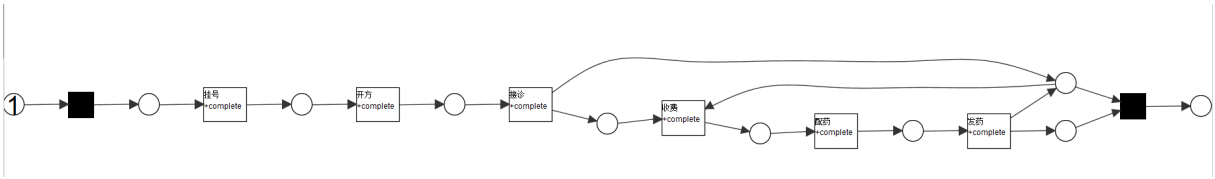


转换为Petri网如下图：



2.4 ILP-based流程挖掘算法

ILP-based（基于约束逻辑编程的）流程挖掘算法是一种通过逻辑约束和推理来发现和分析业务流程的算法， 这种方法强调通过逻辑推理与约束的组合，从实际数据中自动生成符合实际业务过程的流程模型。下面是得到的模型:



3 模型评估

下面基于**拟合度(Fitness)**，**泛化度（Generalization）**，**精确度(Precision)**，**简洁度(Simplicity)**四个建模质量指标评估各个模型，以**高、中、低**为量化标准。

3.1 Alpha算法

可以看到Alpha算法得出的模型要么结构过于独立无法使用，要么过于复杂，泛化度较差。因此拟合度与精确度高，但泛化度和简洁度低。

3.2 Fuzzy算法

Fuzzy算法适用于处理复杂且模糊的流程数据，对于含噪音的数据有奇效。可以看到得到的模型拟合度一般、泛化度与简洁度高，但在模型中存在显著错误（报告审核在报告发布之后），推测是因为数据中报告相关的事件发生时间一样导致的，所以精确度低。

3.3 启发式算法

启发式算法适合噪音多，日志不完整的流程数据，通过启发式规则补充流程数据缺少的信息。可以看到启发式算法得到的模型简洁度与泛化度都不错，拟合度上也还行，但是也存在显著错误（报告审核与检查没有依赖关系），所以精确度一般。

3.4 ILP-based算法

ILP-based的流程挖掘算法基于归纳逻辑推理，给出的模型的简洁度极高，但是精确度、拟合度极低，泛化度一般。

3.5 人工建模

具体模型见下一节。人工建模通过观察流程数据并结合生活经验进行处理，对于拟合度和精确度都有不错的表现，简洁度也可以，但是泛化度较低。

最终总结如下表:

算法	Fitness	Generalization	Precision	Simplicity
Alpha	高	低	高	低
Fuzzy	中	高	低	高
启发式	高	高	低	高
ILP-based	低	中	低	高
人工建模	高	低	高	高

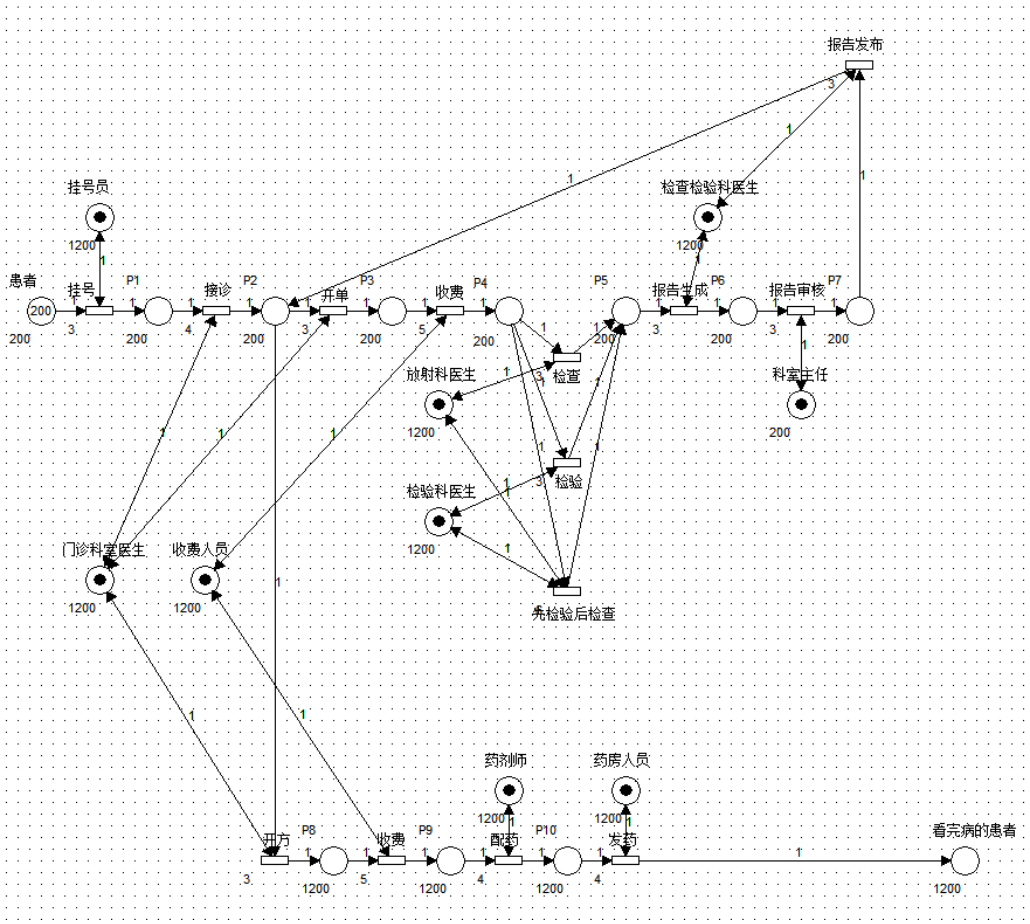
Table 1: 不同算法的模型评估结果

可以看出对于此医院流程数据，**启发式算法**与**人工建模**是比较适合的流程挖掘方法。

4 Petri网模型构建与仿真

4.1 模型构建

根据流程挖掘结果，参考实际生活经验，建立Petri网如下：



4.2 模型仿真

由于Petri网的转移是不可并发的（即在指定的延迟内，只能转移一次托肯），因此无法模拟多个医院人员同时处理患者的场景。不妨配置所有人员都为1个（相当于开关变量所以数量不影响），患者初始值**200个**，尝试模拟串行流水线下模型处理患者的速度。

为了得到各个转移的时延，使用python预处理数据时，顺便统计了每个ACTIVITY的平均DURATION，保存到文件中。结果如下：

	A	B
1	ACTIVITY	DURATION
2	发药	258.7987805
3	开单	216.1113689
4	开方	217.4949944
5	报告	230.375
6	报告发布	217.4261682
7	报告审核	214.3164794
8	报告生成	217.2296015
9	挂号	177.569
10	接诊	230.294
11	收费	303.2495935
12	检查	214.9873418
13	检验	213.3440367
14	配药	248.3987805

由于HPSim软件最多只能模拟1000步，需要将时延单位放大来尽可能模拟久一点的时间。在此取分钟为单位，并且向下取整，比如发药的时延就是4分钟，在模型里为4。

模拟10次，得到每一次完成看病的患者数量，平均为**65.6名**，平均结束时间为**513.7（分钟）**，则串行流水线看病的速度约为7.83分钟/个患者。此数据比单个患者看病所花时间要快很多，是因为Petri网中模拟的是流水线行为，不同的事件之间存在并发，但一个转移不可并发处理多个托肯。以大医院一天挂号**1000个**左右为标准，工作时间取上午8:00到12:00，下午13:30到17:00，共**7.5个小时**，**450分钟**，则大约在每个岗位需要配置**17.4个**人员。

上述计算情境并未考虑各个医护人员需要处理的事件多少有所区别，可以简单的认为一个岗位参与多少种事件就需要几倍的人员数，因此最终的建议人员配置为：

挂号员	收费人员	放射科医生
17	35	17
检查检验科医生	检验科医生	科室主任
35	17	17
药剂师	药房人员	门诊科室医生
17	17	52

Table 2: 医护人员配置建议表

由于经典Petri网模型的局限性，此仿真计算结果的准确性可能不甚理想，后续可以进行优化和修改，针对这些问题进行改进，以求达到更接近真实情境的模拟结果。