**人工智能ACMG评级预测**

**技术方案**

|  |  |  |  |
| --- | --- | --- | --- |
| **文档版本号** | V1.0 | **文档编号** |  |
| **文档密级** | 项目组可见 | **归属部门/项目** | ACMG智能化评级 |
| **文档名称** | 人工智能ACMG评级预测技术方案 | **产品名** |  |
| **编写人** | 李雪洁 | **编写日期** | 2022.10.09 |

**修订记录**

备注：文档正文中，蓝色字体为新增/修改部分，红色字体为重点强调部分

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **编号** | **文档版本** | **变化状态** | **修订页码** | **修订说明** | **修订人** | **修订日期** |
| 1 | V1.0 | C | 创建 | 创建 | 李雪洁 | 2022.10.09 |
| 2 |  |  |  |  |  |  |
| 3 |  |  |  |  |  |  |
| 4 |  |  |  |  |  |  |
| 5 |  |  |  |  |  |  |
| 6 |  |  |  |  |  |  |
| 7 |  |  |  |  |  |  |
| 8 |  |  |  |  |  |  |
| 9 |  |  |  |  |  |  |
| 10 |  |  |  |  |  |  |

\*变化状态：C――创建，A——增加，M——修改，D——删除

**目 录**

[一、 背景 1](#_Toc21646)

[二、 任务概述 1](#_Toc21659)

[三、 技术方案 1](#_Toc28921)

[1. 证据评级方案 1](#_Toc22085)

[1.1 PVS1 1](#_Toc1885)

[1.2 PS1 1](#_Toc24719)

[1.3 PS2和PM6 2](#_Toc8511)

[1.4 PS3和BS3 3](#_Toc1357)

[1.5 PS4 3](#_Toc1528)

[1.6 PM1 3](#_Toc24548)

[1.7 PM2 4](#_Toc22277)

[1.8 PM3 4](#_Toc22757)

[1.9 PM4 5](#_Toc31662)

[1.10 PM5 5](#_Toc16259)

[1.11 PP1 5](#_Toc16164)

[1.12 PP2 6](#_Toc12243)

[1.13 PP3和BP4 6](#_Toc23617)

[1.14 PP4 7](#_Toc13302)

[1.15 BA1 8](#_Toc19596)

[1.16 BP1 8](#_Toc16281)

[1.17 BP2 8](#_Toc7225)

[1.18 BP3 8](#_Toc24688)

[1.19 BP7 9](#_Toc21301)

[1.20 BS1、BS2、BS4、BP5 9](#_Toc5017)

[2. 证据评级流程图 9](#_Toc17278)

[3. 致病性评级方案 10](#_Toc28429)

[四、 技术进度计划表 11](#_Toc20313)

1. **背景**

下一代测序（NGS）技术已经深刻影响了基因组研究进程，使得我们可以在较短时间内同时查询到数千个基因变异的目标区域，发现了更多的基因-疾病关联性，增加了我们对从遗传学到肿瘤学和微生物学的各个医学领域的理解。NGS往往会产出大量的数据结果，基于NGS检测结果解释每个个体产生的成百上千个变异的临床意义是一个巨大的挑战。

美国医学遗传学和基因组学学院（ACMG）和分子病理学协会（AMP）修订并更新了指南来解释这些变异。指南建议使用五个级别（致病性、致病性、不确定意义，可能良性和良性）来评判变异的类型。指南整合了可用的信息（即人群频率、计算和预测数据、功能数据、分离数据和其他），提供了26个评级指标。这份指南在临床上具有巨大的意义。

目前，临床上对于遗传病的变异评级是结合许多的预测工具结果、数据库信息、文献报道等完成的。没有某种工具可以满足ACMG指南中的所有标准，在这种情况下，必须开发新的方法来改进评级系统的颗粒度，从而为分类带来更多的确定性。

1. **任务概述**

机器学习方法，从Logistic回归到基于树的技术，在过去的几年中已经被应用于区分致病性变异和良性变异。我们将利用机器学习等NLP算法对变异位点进行致病性评级，提高自动化评级的准确性，减少人工评级的时间。

1. **技术方案**

ACMG指南赋予了致病性评级明确的方向，基于此指南，我们深度剖析了26种证据的特征，将其分为基于规则实现的12种证据和基于文献阅读获得的14种证据。针对于规则证据，利用既定阈值及部分数据库结果自动化判断证据类型；针对于文献证据，利用NLP算法中的关系抽取、文本分类、文本向量化、图像识别等模型自动化判断证据类型。将证据评级结果作为主要特征，利用机器学习算法，达到智能化致病性评级的目的。

1. 证据评级方案

1.1 PVS1

**证据描述**

当一个疾病的致病机制为功能丧失(LOF)时, 无功能变异(无义突变、移码突变、经典±1 或2 的剪接突变、起始密码子变异、单个或多个外显子缺失)。

**技术方案**

使用华大PVS1开源代码进行规则改善，对符合条件的目标位点给与PVS1。

### 1.2 PS1

**证据描述**

该变异与先前已确定为致病性的变异具有相同的氨基酸改变，用于“不同的”碱基改变引起的相同氨基酸改变的错义变异致病性评价，例如：NM\_000271.5(NPC1):c.2974G>A (p.Gly992Arg)和NM\_000271.5(NPC1):c.2974G>C (p.Gly992Arg)。

**技术方案**

1. 已知的致病性位点：

Step1 提取ClinVar数据库收录的P/LP的位点、HGMD收录P的位点，然后进行注释，构建氨基酸变化的数据库；

Step2 用vep等注释工具对输入变异注释出氨基酸变化信息；

Step3 将目标变异氨基酸信息在本地数据库中比对氨基酸变化，若比对成功给与PS1证据。

1. 未知的致病性位点：在文献中定位到目标氨基酸，摘取定位位置的上下文，进行氨基酸与变异位点的关系抽取。对关系中的变异位点进行致病性分类，若与目标位点不同且变异致病，则提示目标位点为PS1。

针对文本描述具体评级步骤：

Step1 判断输入变异是否是错义变异，若是则转化变异检索式搜索相关文献：针对PS1变异变化情况，基于正则搜索到相同氨基酸改变的变异相关文献；

Step2 定位文献变异相关语句，利用关系抽取模型抽取基因、位点、氨基酸关系；若氨基酸与输入变异氨基酸相同，但位点与输入变异位点不同，则进行下一步骤；

Step3 利用NLP分类模型判断定位到的变异是否致病，若变异为致病，则评为PS1。

1.3 PS2和PM6

**证据描述**

PS2：新发位点“de novo”， Trios家系或者其他结果验证了亲缘关系。

PM6：新发位点“de novo”， 无结果验证亲缘关系。

**技术方案**

1. 患者有家系验证：生信根据患者及家系验证结果，判断是否为新发位点，如果是，给与PS2。
2. 患者无家系验证：在文献中定位目标位点，摘取定位位置的上下文，进行关系抽取，需要识别的关系包括位点与新发的关系、位点与家系验证的关系。针对文本描述具体评级步骤：

Step1 文献定位到目标变异的上下文，所匹配片段含“de novo”关键词则进行下一步骤；

Step2 将新发变异看作一个变异新发事件，运用事件抽取模型，抽取其中的基因、转录本、核苷酸、氨基酸、新发、家系验证6个要素；

Step3 若抽取要素含有基因(或转录本)、核苷酸（或氨基酸）、新发、家系验证4类要素则评为PS2；若抽取要素含有基因(或转录本)、核苷酸（或氨基酸）、新发3类要素则评为PM6。

1.4 PS3和BS3

**证据描述**

PS3：体内和体外功能实验已明确变异会导致基因功能受损；

BS3：体内和体外功能实验已明确变异不会导致基因功能受损。

**技术方案**

1. 提取文献摘要、正文、图表描述等文字内容，并以句为单位进行切分。在文献中定位上一步中生成的所有关键词。
2. 提取关键词所在句子及其前后文，分析每一个文本片段，识别句子中的实体并判断实体之间的关系。需要识别的实体包含：实验名称、实验结果、突变名称。需识别的关系包括：实验名称与实验结果的关系、实验结果与突变名称的关系。
3. 使用自定义规则或分类模型对实验结果文本进行分类，判断该变异是否导致基因功能受损；使用自定义规则或分类模型对实验名称进行分类，判断该实验是否在预设的实验范围内。
4. 当文中检测到对应的变异名称，在上下文中检测到针对该变异的体内外实验，且实验方法符合要求、实验结果为有影响时，可给出PS3；无影响时，给出BS3。

1.5 PS4

**证据描述**

主要考虑常染色体显性遗传类型，在满足PM2的条件下再进行PS4的评级(变异是否满足PM2生信会给出)。统计位点相关文献的患病群体和对照群体，且患病群体>对照群体。

**技术方案**

1. 文本数据：

Step 1 根据基因和变异检索出相关文献；

Step 2 根据关键字，定位到上下文句子；

Step 3 使用关系抽取模型判断出基因-氨基酸-核苷酸之间的关系，同时判断出该文本是否描述关联变异的患病人群个数，提取其中的数字，用于最终总个数的统计。

1. 图像数据（家系图）：**第二阶段实现**

Step 1 根据基因和变异检索出相关文献；

Step 2 根据关键字找到关联的家系图；

Step 3 使用图像检测和分类模型，获取其中患者个数信息。

1. 文献表格：**第二阶段实现**

Step 1 根据基因和变异检索出相关文献；

Step 2 返回该基因/变异关于患者情况描述的整体表格，提供给专家进行判断。

1.6 PM1

**证据描述**

该证据用于错义变异分类，引起蛋白截断的变异和同义突变不可用；位于热点突变区域。

**技术方案**

Step1 用vep等注释工具对输入变异注释出变异分类及氨基酸变化信息；

Step2 根据注释的结果排除同义突变和无义突变；

Step3 提取ClinVar数据库收录的P/LP的位点，并使用VEP进行注释，构建本地数据库；

Step4 根据本地数据库统计目标变异上下游10个氨基酸范围内的致病变异数量。

1.7 PM2

**证据描述**

ESP数据库、千人数据库、EXAC数据库中正常对照人群中未发现的变异（或隐性遗传中极低频位点）。

**技术方案**

引用gnomAD数据库结果(EAS和All类别均满足)：AR <= 0.005;AD<= 0.00005即为PM2，其他人群频率暂不考虑。

Step1 使用VEP对目标变异注释gnomAD人群频率，包括所有人群和EAS；

Step2 根据变异所在基因，注释该基因对应疾病（OMIM）的遗传模式；

Step3 根据不同的遗传模式，对目标变异进行PM2结果预测。

1.8 PM3

**证据描述**

隐性遗传

**技术方案**

1. 有家系样本：根据OMIM数据库对目标变异所在基因的疾病遗传模式进行注释。
2. 无家系样本：

Step1 文字数据：提取关键词所在句子及其前后文。分析每一个文本片段，识别句子中的实体并判断实体之间的关系。需要识别的实体包含：纯合杂合状态，突变名称，疾病名称。需识别的关系包括：纯合杂合状态与突变名称之间的关系，疾病名称与突变名称之间的关系。给出PM3证据需同时满足以下两个条件：

1. 文中检测到对应的变异名称，在上下文中检测到“复合杂合”或“纯合”，并确认该状态描述对象为当前检测的变异；
2. 在上下文中检测到疾病名称，并确认当前突变导致该疾病。

Step2 图像数据（家系图）：解构家系图的成员关系以及患病情况，判断该变异是否为隐性遗传。**第二阶段实现**

Step3 文献表格：定位到表格中的目标位点，判断其是否存在复合杂合，隐性纯合状态。**第二阶段实现**

1.9 PM4

**证据描述**

非重复区框内插入/缺失或终止密码子丧失导致的蛋白质长度变化。

**技术方案**

Step1 使用hg19\_rmsk数据库注释出重复区域信息，这里的重复区域包括短/长串联、散布式重复及低复杂度区域等；

Step2 考虑蛋白质长度是否发生改变，即是否存在框内插入和缺失；

Step3 如在非重复区域出现蛋白质长度改变，则标记PM4；

Step4 如出现stop\_lost，则标记PM4。

1.10 PM5

**证据描述**

新的错义突变导致氨基酸改变,之前在同一位点导致改变为另一种氨基酸的变异已经证实是致病的。

**技术方案**

1. 已知的致病性位点：构建本地库（引用clinvar/hgmd数据库），将目标变异的氨基酸位置与本地库进行比较，与已知致病性变异的氨基酸位置相同，评为PM5。
2. 未知的致病性位点：查找同一氨基酸位置上氨基酸变化不同的变异文献，摘取定位与该氨基酸的上下文，对该氨基酸变化进行致病性分类，若变异致病则提示目标位点为PM5。

针对文本描述具体评级步骤：

Step1 判断输入变异是否是错义变异，若是则转化变异检索式搜索相关文献：针对PM5变异变化情况，基于正则搜索到相同氨基酸位置上氨基酸变化不同的相关文献；

Step2 定位文献变异相关语句，利用NLP分类模型判断定位到的变异是否致病，若变异为致病，则评为PM5。

1.11 PP1

**第二阶段实现**

**证据描述**

家系具有共分离现象，在满足PM2情况下使用PP1证据。

PP1：对于隐性，突变与疾病表型在一个患病的亲属中检测到此变异，和对于显性，突变与疾病表型在两个患病的亲属中检测到此变异；

PP1\_Moderate：对于隐性，突变与疾病表型在两个患病的亲属中检测到此变异，和对于显性，突变与疾病表型在四个患病的亲属中检测到此变异。

PP1\_Strong：对于隐性，突变与疾病表型在三个患病的亲属中检测到此变异，和对于显性，突变与疾病表型在五个患病的亲属中检测到此变异。

**技术方案**

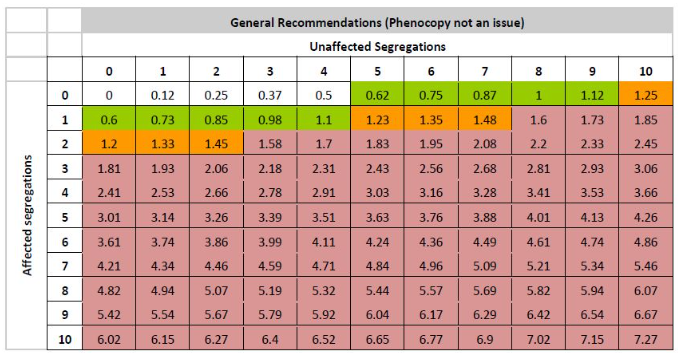
1. 文本数据：

Step 1 根据基因和变异检索出相关文献

Step 2 根据关键字，定位到上下文句子

Step 3 使用关系抽取模型判断出基因-氨基酸-核苷酸之间的关系，同时判断出该文本是否描述共分离的情况。

1. 图像数据：对文献中的家系图进行解构，统计出患病个数与未患病个数。利用下表规则计算Z值。
2. 显性遗传，Z值（LOD）大于等于0.6为PP1，大于等于1.2可以上调到中等致病性证据PP1\_M，大于等于1.5可以上调到强致病性证据PP1\_S；
3. 隐性遗传病的家系，父母一般为携带者，在计算共分离的次数时只计算先证者的兄弟姐妹。



1.12 PP2

**证据描述**

某个基因的错义变异是造成某种疾病的发生，并且这个基因中良性变异所占的比例很小，在这样的基因中所发现的错义变异。

**技术方案**

1. 在gnomAD官网下载gnomad.v2.1.1.lof\_metrics.by\_gene.txt.gz文件，截取gene，syn\_z，mis\_z，lof\_z四列作为插件注释文件gnomad\_Z\_values.txt；
2. 仿造vep自带的插件模块制作Plugins 名为gnomadZvalue.pm模块；
3. 用vep注释gnomadZvalue Plugins，主要根据mis\_z值是否大于等于3.09作为PP2判断依据。

1.13 PP3和BP4

**证据描述**

PP3：多种统计方法预测出该变异会对基因或基因产物造成有害的影响，包括保守性预测、进化预测、剪接位点影响；对于错义突变，根据SVI的判定标准，使用meta-predictor tool (REVEL)代替多种预测工具。

BP4：多种统计方法预测出该变异会对基因或基因产物无影响。

**技术方案**

1. 错义变异：

Step 1 根据VEP注释结果判定变异是否为错义突变；

Step 2 如其REVEL\_score在0.7至1之间则标记PP3；

Step 3 如其REVEL\_score在0至0.4之间则标记BP4。

1. 剪切突变：

Step 1 VEP注释：使用插件，注释包括 SpliceAI、MaxEntScan、dbscSNV的剪接预测

Step 2 判断dbscSNV、SpliceAI、MaxEntScan剪接结果预测：

1. dbscSNV：

如果ada\_score > 0.6 或者 rf\_score > 0.6: pp3\_t1 = 1

1. SpliceAI【默认突变位点和剪切位点最大距离为50】：

如果max(DS\_AG, DS\_AL, DS\_DG, DS\_DL) > 0.5: pp3\_t2 = 1

1. MaxEntScan：

如果(MaxEntScan\_diff/MaxEntScan\_ref) > 15%: pp3\_t3 = 1

PS: 仅用于评价SNVs within the native splice site的splice site loss。

Indel within the native splice site的splice site loss， 以及de novo splice site gain需要参考MES-SWA。阈值标准未能找到文献依据，暂按照15%

Step 3 PP3判别：

如果pp3\_t1 == 1 或者 pp3\_t2 = =1 或者 pp3\_t3 == 1:

则标记为PP3

1.14 PP4

**证据描述**

1. 需要考虑表型的特异性。如果患者的表型具有特异性，同时携带导致该表型的变异点，则直接给出PP4的评分。如果患者无相关表型，需要从文献中查找，若文献中记录的携带该变异的患者具有特异表型，则给出PP4评分。
2. 表型特异的定义和单基因遗传病有关，目前未进行统计，靠人为主观判断，因此，通过文献给出PP4评级暂不考虑，后续第二阶段加入，同时，遗传病会对单基因病进行汇总，明确对表型特异的定义

**技术方案**

step 1: 从临床信息中提取出患者的所有表型；

step 2: 计算患者表型和疾病描述之间的匹配度(由疾病关联到基因)；

step 3: 根据匹配分值(双向)，判断是否特异；

说明：步骤中的双向表示既考虑患者的表型特异，又考虑疾病描述的特异性。

1.15 BA1

**证据描述**

MAF>5%

**技术方案**

引用人群频率gnomAD数据库结果（暂时只看All），对变异位点进行注释，若MAF>5%，即为BA1。

1.16 BP1

**证据描述**

已知一个疾病的致病原因是由于某基因的截短变异，在此基因中所发现的错义变异。

**技术方案**

1. InterVar方法：首先对ClinVar数据库进行清洗，去除未被MedGen注释的变异以及人群频率>0.05的变异（ExAC\_EAS，1000g2015aug\_eas，gnomAD\_exome\_EAS）；对于特定的基因，如果80%且至少一个变异以上的致病变异都是截断变异，那么这个基因发生的错义变异则给出一个BP1的证据；
2. VarSome方法：对于特定的基因，在所有的non-VUS的错义变异中，91.57%的错义变异为良性则这个基因发生的错义变异给出一个BP1的证据；
3. 数据库方面初步打算整合ClinVar和HGMD数据库，上述两个方法一个从致病变异的角度，一个从错义变异的角度，两者可以结合。

1.17 BP2

**第二阶段实现**

**证据描述**

显性（反式）/隐性（顺式）。

**技术方案**

表格数据：解构表格中的位点信息，判断其顺反式，给出BP2证据。目前图表解析较难实现，第二阶段在开发。

1.18 BP3

**证据描述**

功能未知重复区域内的缺失/插入, 同时没有导致基因编码框改变。

**技术方案**

1. 使用hg19\_rmsk数据库注释出重复区域信息，这里的重复区域包括短/长串联、散布式重复及低复杂度区域等；
2. 考虑蛋白质长度是否发生改变，即是否存在框内插入和缺失；
3. 如在重复区域出现蛋白质长度改变，则标记BP3。

1.19 BP7

**证据描述**

同义突变且不影响剪接。

**技术方案**

1. 根据VEP注释结果判定变异是否为同义变异：

Consequence注释字段匹配上"synon" or "coding-synon"，但不匹配"nonsynon"。

1. 判断dbscSNV、SpliceAI、MaxEntScan剪接结果预测：
2. dbscSNV

如果ada\_score < 0.6 且 rf\_score < 0.6: BP7\_s1=1

如果ada\_score = “.” 且 rf\_score =“.”: BP7\_s1=1

[参考InterVar: absent means it is not in the splice consensus sequence]

1. SpliceAI【默认突变位点和剪切位点最大距离为50】：

如果max(DS\_AG, DS\_AL, DS\_DG, DS\_DL) < 0.5： BP7\_s2=1

如果DS\_AG, DS\_AL, DS\_DG, DS\_DL为”.”: BP7\_s2=1(参考dbscSNV)

1. MaxEntScan

如果(MaxEntScan\_diff/MaxEntScan\_ref) < 15%：BP7\_s3=1

PS: 仅用于评价SNVs within the native splice site的splice site loss。

Indel within the native splice site的splice site loss， 以及de novo splice site gain需参考MES-SWA。阈值标准未能找到文献依据，暂按照15%。

1. BP7判别：

如果变异为同义突变，且BP7\_s1=1 and BP7\_s2=1 and BP7\_s3=1，则标记BP7。

1.20 BS1、BS2、BS4、BP5

**证据描述**

BS1：MAF>患病率

BS2：在健康成年人中存在的位点

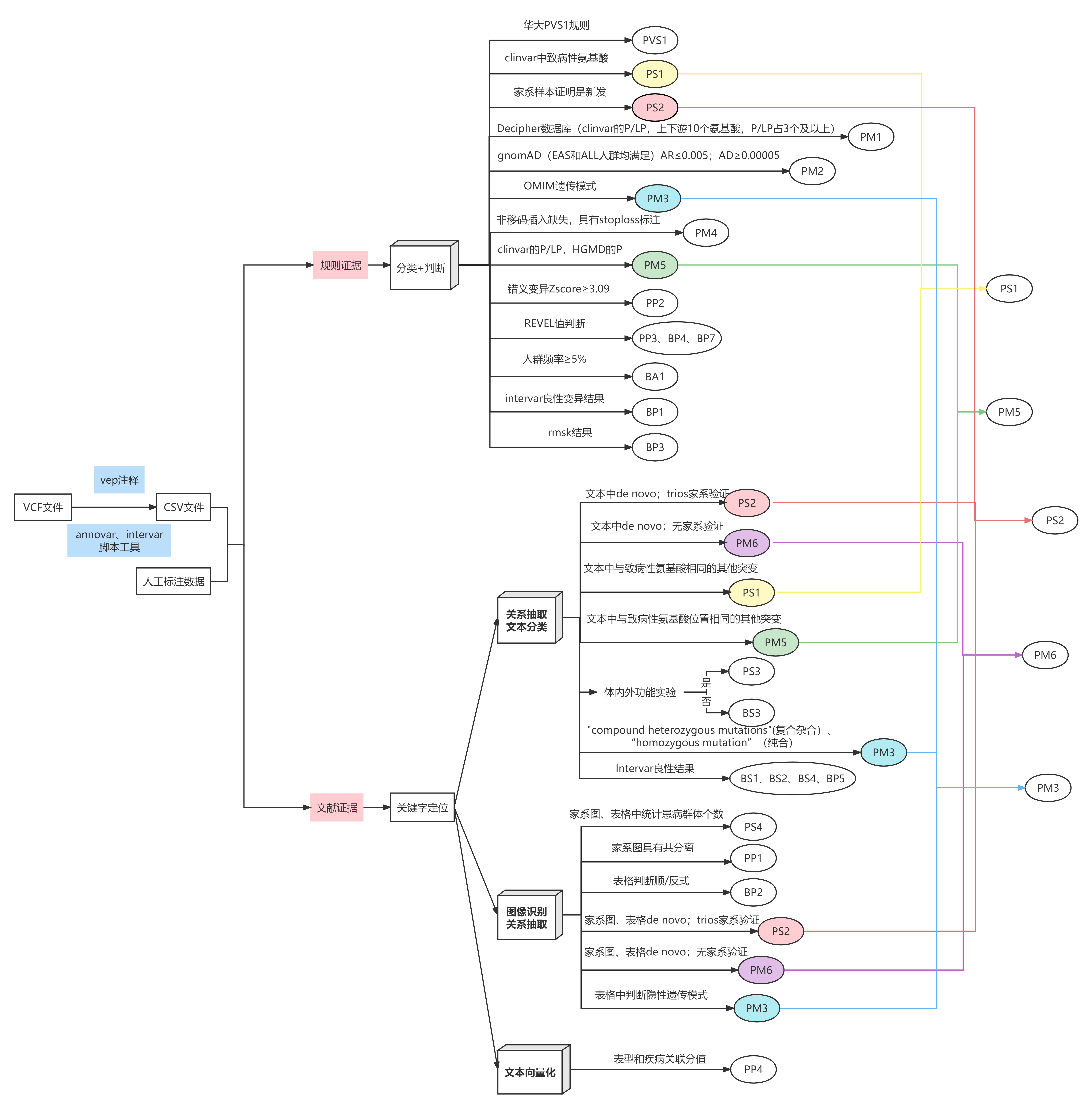
BS4：缺乏共分离

BP5：其他已知病例中变异

**技术方案**

目前先引用Intervar的结果，**或者第二阶段再开发**。

1. 证据评级流程图



1. 致病性评级方案

借助机器学习（eg:LR,RF）在分类和优先级排序任务上的出色能力，基于变异样本的ACMG评级信息及通过基因注释工具（eg:eVal,VEP）生成的注释信息，构建自动化的机器学习模型，以使其在未知的变异样本上具有高准确的致病性/良性预测能力，为基因诊断致病性分析提供辅助决策依据。

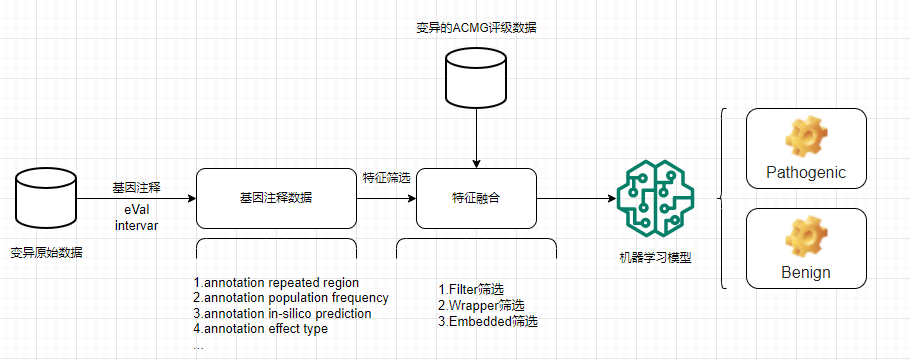
**技术方案**

1. 特征构造：基于变异样本元数据进行基因注释，生成后续分类排序模型的数据基础。包括以下方面数据。
2. annotation repeated region
3. annotation population frequency
4. annotation in-silico prediction
5. annotation effect type

...

1. 特征筛选：运用特征筛选方案，提取高稳定性，高区分度的特征。包括以下方法。
2. Filter筛选
3. Wrapper筛选
4. Embedded筛选

**第二阶段开发**



1. 技术进度计划表

|  |  |  |  |
| --- | --- | --- | --- |
| 任务 | 描述 | 开始时间 | 完成时间 |
| 证据预测开发 | 对ACMG的26条证据进行自动化评级。 | 2022.10.14 | 2022.10.31 |
|  |  |  |  |
|  |  |  |  |