## 理赔表的数据预处理

在理赔表中，首先我们对缺失值进行了检查，未发现存在缺失值。之后，我们对异常值进行筛查，箱型图

最后，我们对重点的死因数据进行检查。

例如ID为2142913的数据中，Reason2为’C16.0:胃恶性肿瘤’，然而标准的ICD-10编码则为’C16.0:贲门癌’，此类错误会极大的影响之后的自然语言处理，因此爬取了ICD-10数据，对Reason2字段进行了校准。

166784中，Reason2为’M8000/6:骨质疏松伴有病理性骨折’，然而这是ICD-O

通过对数据的粗览，我们注意到Reason2通常来看是是导致死亡的直接原因，但正因此使得Reason1和Reason2的描述存在部分矛盾的情况。例如，ID为1900712的数据中，Reason1给出的结果是'客户是自然死亡 ;。'，而Reason2则给出的却是’ A27.0:钩端螺旋体病'。这类矛盾的数据会影响我们对死因的归纳，因此首先我们需要给出Reason1字段和Reason2字段之间的逻辑关系。

为了完成这一目标，我们考虑转化为自然语言推理任务（Natural Language Inference，NLI）。该任务负责判断一个前提（premise）是否能够推导出一个假设（hypothesis），并给出 entailment、contradiction 或者 neutral 的分类结果。在上述的例子中，若把'丙型肝炎'作为一个假设，心脏停搏作为一个前提，那么模型应给出contradiction的分类结果。

我们使用了基于xlm-roberta-large在NLI数据集进行finetune过的xlm-roberta-large-xnli模型（来自论文：Unsupervised Cross-lingual Representation Learning at Scale），更适用于当前任务。其中，由于Reason2字段采用了ICD-10编码，因此其标准程度更高，适合作为前提（模型所训练的NLI数据集中，前提通常是一些标准的描述）。

因此，使用该模型推理的pipeline如下所示：

分词器f输入premise： Reason2字段的数据和hypothesis：Reason1字段的数据，从而将文本转化为词向量X，进而输入到模型g中，得到一个logit分布L，最后分类结果取logit分布L中最大者即可

最后，我们得到了如下结果：