(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 113271849 A
(43)申请公布日 2021.08.17

(21)申请号 201980078901.3
(22)申请日 2019.11.21
(30)优先权数据
62/773,028 2018.11.29 US
62/783,733 2018.12.21 US
(85)PCT国际申请进入国家阶段日
2021.05.28
(86)PCT国际申请的申请数据
PCT/US2019/062561 2019.11.21
(87)PCT国际申请的公布数据
WO2020/112478 EN 2020.06.04
(71)申请人 私募蛋白质体公司
地址 美国科罗拉多州
(72)发明人 Y·夏甲 G·达塔 L·亚历山大 M·欣特伯格
(74)专利代理机构 北京汇知杰知识产权代理有限公司 11587
代理人 杨巍 柴春玲
(51)Int.Cl.
A61B 5/117(2016.01)
A61B 5/00(2006.01)

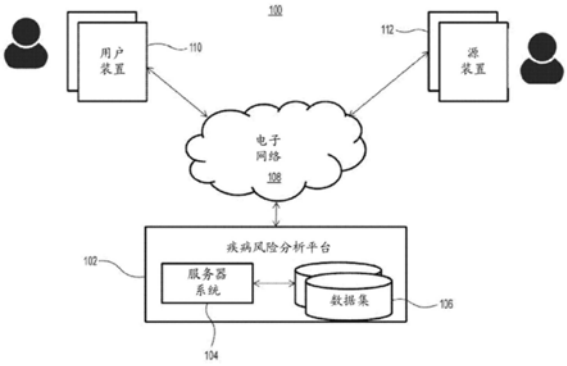
权利要求书4页 说明书14页 附图7页

(54)发明名称

结合类别不平衡集降采样与生存分析的疾病风险确定方法

(57)摘要

一种用于利用生存分析对类别不平衡集进行降采样的方法，其包括：获取类别不平衡数据集，其中所述类别不平衡数据集包括来自多个受试者的生物数据，其中每个受试者的所述生物数据包括观察结果、时间值和多个临床测量结果，并且其中所述生物数据被归类为多数数据类别或少数数据类别的一部分，其中所述多数数据类别具有数量比所述少数数据类别更多的观察结果；对所述类别不平衡数据集进行降采样，其中所述降采样导致所述多数数据类别具有数量与所述少数数据类别相等或基本相等的观察结果；以及利用生存分析对所述降采样后的数据集进行交叉验证以生成生存模型；其中所述观察结果包括在特定时间值处的事件或无事件。



CN 113271849 A

1.一种方法,其包括:

a)获取类别不平衡数据集,其中所述类别不平衡数据集包括来自多个受试者的生物数据,其中每个受试者的所述生物数据包括观察结果、时间值和多个临床测量结果,并且其中所述生物数据被归类为多数数据类别或少数数据类别的一部分,其中所述多数数据类别具有数量比所述少数数据类别更多的观察结果;

b)对所述类别不平衡数据集进行降采样以生成降采样后的数据集,其中所述降采样导致所述多数数据类别具有数量与所述少数数据类别相等或基本相等的观察结果;以及

c)利用生存分析对所述降采样后的数据集进行交叉验证以生成生存模型;

其中所述观察结果包括在特定时间值处的事件或无事件。

2.如权利要求1所述的方法,其中所述生存模型的AUC、敏感性、特异性和/或C指数比其中在所述生存分析之前未对所述类别不平衡数据集进行降采样的生存模型的AUC、敏感性、特异性和/或C指数更接近1。

3.如权利要求1所述的方法,其中所述类别不平衡数据集是生存数据集。

4.如权利要求1所述的方法,其中所述事件是受试者的疾病、病症或病状。

5.如权利要求1所述的方法,其中所述生存分析选自由Cox比例风险分析、随机森林分析、加速失效时间分析及其任何组合组成的组。

6.如权利要求5所述的方法,其还包括弹性网罚分。

7.如权利要求1所述的方法,其中所述交叉验证是至少2折、3折、4折、5折、6折、7折、8折、9折、10折、11折、12折、13折、14折、15折、16折、17折、18折、19折、或20折交叉验证。

8.如权利要求1所述的方法,其中所述生存模型包括5至1,000个特征,其中每个特征选自由蛋白测量结果、临床因素及其组合组成的组。

9.如权利要求8所述的方法,其中所述临床因素选自由年龄、体重、血压、身高、BMI、胆固醇、性别及其组合组成的组。

10.如权利要求1所述的方法,其中所述临床测量结果选自蛋白质组学测量结果、基因组测量结果、转录组测量结果、代谢组学测量及其组合。

11.如权利要求1所述的方法,其中所述交叉验证选自k折交叉验证、蒙特卡洛交叉验证和排除N验证。

12.如权利要求1所述的方法,其中所述多数数据类别是所述类别不平衡数据集的95%,并且所述少数数据类别是所述类别不平衡数据集的5%。

13.如权利要求1所述的方法,其中所述多数数据类别是所述类别不平衡数据集的90%,并且所述少数数据类别是所述类别不平衡数据集的10%。

14.如权利要求1所述的方法,其中所述多数数据类别是所述类别不平衡数据集的85%,并且所述少数数据类别是所述类别不平衡数据集的15%。

15.如权利要求1所述的方法,其中所述多数数据类别是所述类别不平衡数据集的80%,并且所述少数数据类别是所述类别不平衡数据集的20%。

16.如权利要求1所述的方法,其中所述多数数据类别是所述类别不平衡数据集的75%,并且所述少数数据类别是所述类别不平衡数据集的25%。

17.如权利要求1所述的方法,其中所述多数数据类别是所述类别不平衡数据集的70%,并且所述少数数据类别是所述类别不平衡数据集的30%。

18.如权利要求1所述的方法，其中所述多数数据类别是所述类别不平衡数据集的65％，并且所述少数数据类别是所述类别不平衡数据集的35％。

19.如权利要求1所述的方法，其中所述多数数据类别是所述类别不平衡数据集的60％，并且所述少数数据类别是所述类别不平衡数据集的40％。

20.一种方法，其包括：

a)对类别不平衡数据集进行降采样以生成降采样后的数据集，其中所述降采样导致多数数据类别具有数量与少数数据类别相等或基本相等的观察结果；以及

b)利用生存分析对所述降采样后的数据集进行交叉验证以生成生存模型；

其中所述观察结果包括在特定时间值处的事件或无事件；并且

其中所述类别不平衡数据集包括来自多个受试者的生物数据，其中每个受试者的所述生物数据包括观察结果、时间值和多个蛋白测量结果，并且其中所述生物数据被归类为所述多数数据类别或所述少数数据类别的一部分，其中所述多数数据类别具有数量比所述少数数据类别更多的观察结果。

21.如权利要求20所述的方法，其中所述生存模型的AUC、敏感性、特异性和/或C指数比其中在所述生存分析之前未对所述类别不平衡数据集进行降采样的生存模型的AUC、敏感性、特异性和/或C指数更接近1。

22.如权利要求21所述的方法，其中所述AUC是基于确定受试者是否将在指定时间点以前发生事件来计算。

23.一种用于确定疾病风险的计算机实现的方法，其包括：

a)获取类别不平衡数据集，其中所述类别不平衡数据集包括来自多个受试者的生物数据，其中每个受试者的所述生物数据包括观察结果、时间值和多个临床测量结果，并且其中所述生物数据被归类为多数数据类别或少数数据类别的一部分，其中所述多数数据类别具有数量比所述少数数据类别更多的观察结果；

b)对所述类别不平衡数据集进行降采样以生成降采样后的数据集，其中所述降采样导致所述多数数据类别具有数量与所述少数数据类别相等或基本相等的观察结果；以及

c)利用生存分析对所述降采样后的数据集进行交叉验证以生成生存模型；

其中，所述观察结果包括在特定时间值处的事件或无事件；并且

其中步骤b)和步骤c)是利用计算机系统来计算。

24.如权利要求23所述的方法，其中所述生存模型的AUC、敏感性、特异性和/或C指数比其中在所述生存分析之前未对所述类别不平衡数据集进行降采样的生存模型的AUC、敏感性、特异性和/或C指数更接近1。

25.一种可由计算机读取的程序存储装置，所述程序存储装置有形地体现指令程序，所述指令程序可由所述计算机执行以进行用于确定疾病风险的方法的方法步骤，所述方法包括：

a)获取类别不平衡数据集，其中所述类别不平衡数据集包括来自多个受试者的生物数据，其中每个受试者的所述生物数据包括观察结果、时间值和多个临床测量结果，并且其中所述生物数据被归类为多数数据类别或少数数据类别的一部分，其中所述多数数据类别具有数量比所述少数数据类别更多的观察结果；

b)对所述类别不平衡数据集进行降采样以生成降采样后的数据集，其中所述降采样导

致所述多数数据类别具有数量与所述少数数据类别相等或基本相等的观察结果;以及

c)利用生存分析对所述降采样后的数据集进行交叉验证以生成生存模型;

其中所述观察结果包括在特定时间值处的事件或无事件。

26.如权利要求25所述的方法,其中所述生存模型的AUC、敏感性、特异性和/或C指数比其中在所述生存分析之前未对所述类别不平衡数据集进行降采样的生存模型的AUC、敏感性、特异性和/或C指数更接近1。

27.一种用于确定疾病风险的计算系统,其包括:用于存储编程指令的存储器;经配置来执行所述编程指令以进行操作的处理器,所述操作包括:

a)获取类别不平衡数据集,其中所述类别不平衡数据集包括来自多个受试者的生物数据,其中每个受试者的所述生物数据包括观察结果、时间值和多个临床测量结果,并且其中所述生物数据被归类为多数数据类别或少数数据类别的一部分,其中所述多数数据类别具有数量比所述少数数据类别更多的观察结果;

b)对所述类别不平衡数据集进行降采样以生成降采样后的数据集,其中所述降采样导致所述多数数据类别具有数量与所述少数数据类别相等或基本相等的观察结果;以及

c)利用生存分析对所述降采样后的数据集进行交叉验证以生成生存模型;

其中,所述观察结果包括在特定时间值处的事件或无事件。

28.如权利要求27所述的方法,其中所述生存模型的AUC、敏感性、特异性和/或C指数比其中在所述生存分析之前未对所述类别不平衡数据集进行降采样的生存模型的AUC、敏感性、特异性和/或C指数更接近1。

29.一种非暂时性计算机可读介质,其上存储有指令,所述指令可由处理器执行以进行以下操作:

a)获取类别不平衡数据集,其中所述类别不平衡数据集包括来自多个受试者的生物数据,其中每个受试者的所述生物数据包括观察结果、时间值和多个临床测量结果,并且其中所述生物数据被归类为多数数据类别或少数数据类别的一部分,其中所述多数数据类别具有数量比所述少数数据类别更多的观察结果;

b)对所述类别不平衡数据集进行降采样以生成降采样后的数据集,其中所述降采样导致所述多数数据类别具有数量与所述少数数据类别相等或基本相等的观察结果;以及

c)利用生存分析对所述降采样后的数据集进行交叉验证以生成生存模型;

其中所述观察结果包括在特定时间值处的事件或无事件。

30.如权利要求29所述的方法,其中所述生存模型的AUC、敏感性、特异性和/或C指数比其中在所述生存分析之前未对所述类别不平衡数据集进行降采样的生存模型的AUC、敏感性、特异性和/或C指数更接近1。

31.一种用于确定疾病风险的计算机实现的方法,其包括:

a)利用计算机接收类别不平衡数据集,其中所述类别不平衡数据集包括来自多个受试者的生物数据,其中每个受试者的所述生物数据包括观察结果、时间值和多个临床测量结果,并且其中所述生物数据被归类为多数数据类别或少数数据类别的一部分,其中所述多数数据类别具有数量比所述少数数据类别更多的观察结果;

b)利用所述计算机对所述类别不平衡数据集进行降采样以生成降采样后的数据集,其中所述降采样导致所述多数数据类别具有数量与所述少数数据类别相等或基本相等的观

察结果;以及

c)利用所述计算机来利用生存分析对所述降采样后的数据集进行交叉验证以生成生存模型;

其中所述观察结果包括在特定时间值处的事件或无事件。

32.如权利要求31所述的方法,其中所述生存模型的AUC、敏感性、特异性和/或C指数比其中在所述生存分析之前未对所述类别不平衡数据集进行降采样的生存模型的AUC、敏感性、特异性和/或C指数更接近1。

结合类别不平衡集降采样与生存分析的疾病风险确定方法

[0001] 相关申请的交叉引用

[0002] 本申请要求2018年11月29日提交的美国临时专利申请号62/773,028和2018年12月21日提交的美国临时专利申请号62/783,733的优先权,所述临时申请以引用的方式整体并入本文。

技术领域

[0003] 本公开总体上涉及疾病风险确定的领域,并且更具体地,涉及用于处理电子数据以确定疾病风险的系统和方法。

背景技术

[0004] 用于识别与各种疾病相关病状或事件(例如心血管事件、糖尿病诊断、各种癌症类型等)的风险相关联的生物标志物的方法已得到改进,这主要是由于发现了高通量技术,诸如基因测序、转录组学、蛋白质组学和代谢组学。但是,这些技术也因提供表示复杂生物过程的高维数据而使事情复杂化,这可能使提取有意义的生物标志物标签变得困难。

[0005] 当主要目标是正确识别在特定时间段内将经历疾病相关病状或事件的个体时,可以加强通常将仅使用分类方法的分析,实现方式是将分析构造为一种特殊类型的分类问题,它结合分类工具合并了两种生存模型方法。但是,生存分析可能会遭受经历和未经历疾病相关病状或事件的患者的数量之间的不平衡。众所周知,预测分类器通常在不平衡数据上表现不佳,因为模型被“尽可能频繁地”训练为准确的。之所以会发生这种效果,是因为较大的多数类别驱动为模型选择的特征,因为少数类别经常被错误地分类,而多数类别仍然被准确地预测。但是,敏感性和特异性将变得不平衡,使得一者相对于另一者最大化,这取决于哪个组有更多的观察结果。在对健康结果进行建模时,在一个人群中患病率较低是很常见的,这个人群形成少数类别。在这种情况下,特异性将以牺牲灵敏度为代价而最大化,而当目标是要识别尽可能多的面临病状或事件发展的风险的个体时,这将是一个问题。

[0006] 因此,仍然需要替代的方法来改进识别针对特定疾病或病状的分子标签或生物标志物的方式。本公开通过提供用于改进生物标志物发现的方法来满足这样的需求。

发明内容

[0007] 根据本公开的一些方面,所公开的系统和方法涉及对包括时间值的类别不平衡数据集的多数类别(即,具有更多观察结果的类别)进行降采样,以便改进生存分析的敏感性和特异性。降采样的目的是“偏置”分类器,以便它对已诊断和未诊断的个体给予同等关注,以平衡模型的敏感性和特异性。

[0008] 在一个实施方案中,公开了一种方法,所述方法包括:获取类别不平衡数据集,其中类别不平衡数据集包括来自多个受试者的生物数据,其中每个受试者的生物数据包括观察结果、时间值和多个临床测量结果,并且其中生物数据被归类为多数数据类别或少数数据类别的一部分,其中多数数据类别具有数量比少数数据类别更多的观察结果;对类别不

平衡数据集进行降采样以生成降采样后的数据集，其中降采样导致多数数据类别具有数量与少数数据类别相等或基本相等的观察结果；以及利用生存分析对降采样后的数据集进行交叉验证以生成生存模型；其中，所述观察结果包括在特定时间值处的事件或无事件。

[0009] 根据本公开的各方面，所述生存模型的曲线下面积(AUC)、敏感性、特异性和/或C指数比其中在生存分析之前未对类别不平衡数据集进行降采样的生存模型的AUC、敏感性、特异性和/或C指数更接近1。

[0010] 在其他示例中，类别不平衡数据集是生存数据集且/或事件是受试者的疾病、病症或病状。在另外的示例中，生存分析选自由cox比例风险分析、随机森林分析、加速失效时间分析及其任何组合组成的组，包括诸如惩罚回归技术的机器学习适应。方法还可包括弹性网罚分。

[0011] 在其他实施方案中，交叉验证是至少2折、3折、4折、5折、6折、7折、8折、9折、10折、11折、12折、13折、14折、15折、16折、17折、18折、19折、或20折交叉验证。在其他实施方案中，生存模型包括5至1,000个特征，其中每个特征选自由蛋白测量结果、临床因素及其组合组成的组。临床因素选自由年龄、体重、血压、身高、BMI、胆固醇、性别及其组合组成的组。

[0012] 在另外的实施方案中，临床测量结果选自蛋白质组学测量结果、基因组测量结果、转录组测量结果、代谢组学测量及其组合。此外，交叉验证选自k折、广义蒙特卡洛、排除p交叉验证或自举方法。

[0013] 根据本公开的各方面，多数数据类别是类别不平衡数据集的95％并且少数数据类别是类别不平衡数据集的5％，或者多数数据类别是类别不平衡数据集的90％并且少数数据类别是类别不平衡数据集的10％，或者多数数据类别是类别不平衡数据集的85％并且少数数据类别是类别不平衡数据集的15％，或者多数数据类别是类别不平衡数据集的80％并且少数数据类别是类别不平衡数据集的20％，或者多数数据类别是类别不平衡数据集的75％并且少数数据类别是类别不平衡数据集的25％，或者多数数据类别是类别不平衡数据集的70％并且少数数据类别是类别不平衡数据集的30％，或者多数数据类别是类别不平衡数据集的65％并且少数数据类别是类别不平衡数据集的35％，或者多数数据类别是类别不平衡数据集的60％并且少数数据类别是类别不平衡数据集的40％。

[0014] 根据另一个实施方案，公开了一种方法，所述方法包括：对类别不平衡数据集进行降采样以生成降采样后的数据集，其中降采样导致多数数据类别具有数量与少数数据类别相等或基本相等的观察结果；以及利用生存分析对降采样后的数据集进行交叉验证以生成生存模型；其中，观察结果包括在特定时间值处的事件或无事件；其中，类别不平衡数据集包括来自多个受试者的生物数据，其中每个受试者的生物数据包括观察结果、时间值和多个蛋白测量结果，并且其中生物数据被归类为多数数据类别或少数数据类别的一部分，其中多数数据类别具有数量比少数数据类别更多的观察结果。

[0015] 根据本公开的各方面，所述生存模型的AUC、敏感性、特异性和/或C指数比其中在生存分析之前未对类别不平衡数据集进行降采样的生存模型的AUC、敏感性、特异性和/或C指数更接近1。

[0016] 在本公开的示例中，AUC是基于确定受试者是否将在指定时间点以前发生事件来计算。

[0017] 还公开了一种用于确定疾病风险的计算机实现的方法，所述方法包括：获取类别

不平衡数据集，其中类别不平衡数据集包括来自多个受试者的生物数据，其中每个受试者的生物数据包括观察结果、时间值和多个临床测量结果，并且其中生物数据被归类为多数数据类别或少数数据类别的一部分，其中多数数据类别具有数量比少数数据类别更多的观察结果；对类别不平衡数据集进行降采样以生成降采样后的数据集，其中降采样导致多数数据类别具有数量与少数数据类别相等或基本相等的观察结果；以及利用生存分析对降采样后的数据集进行交叉验证以生成生存模型；其中，所述观察结果包括在特定时间值处的事件或无事件；并且降采样及交叉验证步骤是利用计算机系统来计算。

[0018] 根据本公开的各方面，所述生存模型的AUC、敏感性、特异性和/或C指数比其中在生存分析之前未对类别不平衡数据集进行降采样的生存模型的AUC、敏感性、特异性和/或C指数更接近1。

[0019] 还公开了一种可由计算机读取的程序存储装置，所述程序存储装置有形地体现指令程序，所述指令程序可由计算机执行以进行用于确定疾病风险的方法的方法步骤，所述方法包括：获取类别不平衡数据集，其中类别不平衡数据集包括来自多个受试者的生物数据，其中每个受试者的生物数据包括观察结果、时间值和多个临床测量结果，并且其中生物数据被归类为多数数据类别或少数数据类别的一部分，其中多数数据类别具有数量比少数数据类别更多的观察结果；对类别不平衡数据集进行降采样以生成降采样后的数据集，其中降采样导致多数数据类别具有数量与少数数据类别相等或基本相等的观察结果；以及利用生存分析对降采样后的数据集进行交叉验证以生成生存模型；其中，所述观察结果包括在特定时间值处的事件或无事件。

[0020] 根据本公开的各方面，所述生存模型的AUC、敏感性、特异性和/或C指数比其中在生存分析之前未对类别不平衡数据集进行降采样的生存模型的AUC、敏感性、特异性和/或C指数更接近1。

[0021] 还公开了一种用于确定疾病风险的计算系统，所述计算系统包括：用于存储编程指令的存储器，以及经配置来执行编程指令以进行操作的处理器，所述操作包括：获取类别不平衡数据集，其中类别不平衡数据集包括来自多个受试者的生物数据，其中每个受试者的生物数据包括观察结果、时间值和多个临床测量结果，并且其中生物数据被归类为多数数据类别或少数数据类别的一部分，其中多数数据类别具有数量比少数数据类别更多的观察结果；对类别不平衡数据集进行降采样以生成降采样后的数据集，其中降采样导致多数数据类别具有数量与少数数据类别相等或基本相等的观察结果；以及利用生存分析对降采样后的数据集进行交叉验证以生成生存模型；其中，所述观察结果包括在特定时间值处的事件或无事件。

[0022] 根据本公开的各方面，所述生存模型的AUC、敏感性、特异性和/或C指数比其中在生存分析之前未对类别不平衡数据集进行降采样的生存模型的AUC、敏感性、特异性和/或C指数更接近1。

[0023] 还公开了一种非暂时性计算机可读介质，其中所述计算机可读介质上存储有指令，所述指令可由处理器执行以进行以下操作：获取类别不平衡数据集，其中类别不平衡数据集包括来自多个受试者的生物数据，其中每个受试者的生物数据包括观察结果、时间值和多个临床测量结果，并且其中生物数据被归类为多数数据类别或少数数据类别的一部分，其中多数数据类别具有数量比少数数据类别更多的观察结果；对类别不平衡数据集进

行降采样以生成降采样后的数据集,其中降采样导致多数数据类别具有数量与少数数据类别相等或基本相等的观察结果;以及利用生存分析对降采样后的数据集进行交叉验证以生成生存模型;其中,所述观察结果包括在特定时间值处的事件或无事件。

[0024] 根据本公开的各方面,所述生存模型的AUC、敏感性、特异性和/或C指数比其中在生存分析之前未对类别不平衡数据集进行降采样的生存模型的AUC、敏感性、特异性和/或C指数更接近1。

[0025] 还公开了一种用于确定疾病风险的计算机实现的方法,所述方法包括:利用计算机接收类别不平衡数据集,其中类别不平衡数据集包括来自多个受试者的生物数据,其中每个受试者的生物数据包括观察结果、时间值和多个临床测量结果,并且其中生物数据被归类为多数数据类别或少数数据类别的一部分,其中多数数据类别具有数量比少数数据类别更多的观察结果;利用计算机对类别不平衡数据集进行降采样以生成降采样后的数据集,其中降采样导致多数数据类别具有数量与少数数据类别相等或基本相等的观察结果;以及利用计算机来利用生存分析对降采样后的数据集进行交叉验证以生成生存模型;并且其中所述观察结果包括在特定时间值处的事件或无事件。

[0026] 根据本公开的各方面,所述生存模型的AUC、敏感性、特异性和/或C指数比其中在生存分析之前未对类别不平衡数据集进行降采样的生存模型的AUC、敏感性、特异性和/或C指数更接近1。

附图说明

[0027] 图1示出其中可实现本公开的方法、系统和其他方面的网络化计算环境的示例。

[0028] 图2是根据本公开的用于临床数据获取和处理的疾病风险分析平台的高层架构图。

[0029] 图3示出HUNT3 CHD亚群中的心肌梗死(MI)的卡普兰-迈耶生存曲线。

[0030] 图4示出测试集上的MI的卡普兰-迈耶生存曲线,预测事件使所述曲线分层。对于每种方法,使用通过交叉验证来识别的阈值将测试集分为高风险和平均风险的个体。然后针对两个组计算两组卡普兰-迈耶曲线。在逻辑回归模型结果中,每个人都被预测为低风险,因此仅产生一条生存曲线。

[0031] 图5示出测试集上的MI的卡普兰-迈耶生存曲线,其使用降采样后的Cox弹性网模型来预测小于或等于4年的MI。研究了将个体归类为高风险的不同阈值。

具体实施方式

[0032] 除非另外注明,否则技术术语是根据常规用法使用。分子生物学中常见术语的定义可见于Benjamin Lewin,Genes V,Oxford University Press出版,1994(ISBN 0-19-854287-9);Kendrew等人(编辑),The Encyclopedia of Molecular Biology,Blackwell Science Ltd.出版,1994(ISBN 0-632-02182-9);以及Robert A.Meyers(编辑),Molecular Biology and Biotechnology:a Comprehensive Desk Reference,VCH Publishers,Inc.出版,1995(ISBN 1-56081-569-8)。除非另外说明,否则本文所用的所有技术和科技术语均具有如本公开内容所属领域中的普通技术人员通常所理解的相同含义。除非上下文另外清楚地指示,否则单数术语“一个/一种(a/an)”和“所述(the)”包括多个提及物。“包括A或B”

是指包括A、或B、或A和B。应进一步理解，针对核酸或多肽给出的所有碱基大小或氨基酸大小以及所有分子量或分子质量值均为近似值，并且被提供用于描述。

[0033] 此外，本文所提供的范围应理解为关于所述范围内的所有值的速记。例如，1至50的范围应理解为包括来自由1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、21、22、23、24、25、26、27、28、29、30、31、32、33、34、35、36、37、38、39、40、41、42、43、44、45、46、47、48、49或50组成的组的任何数字、数字组合或子范围(以及其分数，除非上下文另外清楚地规定)。除非另外指示，否则任何浓度范围、百分率范围、比率范围或整数范围均应理解为包括在所陈述范围内的任何整数的值，并且适当时包括其分数(诸如整数的十分之一和百分之一)。而且，除非另外指示，否则本文中关于任何物理特征(诸如聚合物亚基、大小或厚度)所陈述的任何数字范围均应理解为包括在所陈述范围内的任何整数。除非另外指示，否则如本文所用，“约”或“基本上由……组成”意指所指示的范围、值或结构的±20％。如本文所用，术语“包括”和”包含”为开放式并且同义地使用。

[0034] 尽管与本文所述的那些类似或相等的方法和材料可在本公开的实践或测试中使用，但下文描述合适的方法和材料。本文提到的所有出版物、专利申请、专利以及其他参考文献均以引用的方式整体并入。在有冲突的情况下，则将以本说明书(包括术语的解释)为准。另外，所述材料、方法和示例仅为说明性的并且不意图为限制性的。

[0035] 如本文所用，“SOMA适配体”或慢解离速率修饰适配体是指具有改进的解离速率特性的适配体。SOMA适配体可使用标题为“Method for Generating Aptamers with Improved Off-Rates”的美国专利号7,947,447中所述的改进的SELEX方法产生。

[0036] 术语“生物样品”、“样品”和“测试样品”在本文中互换地使用，以指代从个体获得或以其他方式得到的任何材料、生物流体、组织或细胞。这包括血液(包括全血、白细胞、外周血单核细胞、血沉棕黄层、血浆以及血清)、痰、眼泪、粘液、鼻洗液、鼻抽吸物、呼吸、尿、精液、唾液、腹膜冲洗物、腹水、囊液、脑膜液、羊水、腺液、淋巴液、乳头抽吸物、支气管抽吸物(例如，支气管肺泡灌洗液)、支气管刷检物、滑液、关节抽吸物、器官分泌物、细胞、细胞提取物以及脑脊髓液。这还包括所有前述项的实验分离的部分。例如，可将血液样品分级成血清、血浆，或分级成含有诸如红血细胞或白血细胞(white blood cell/leukocyte)的特定类型的血细胞的部分。在一些方面，样品可以是来自个体的样品的组合，诸如组织和流体样品的组合。术语“生物样品”还包括含有均质化固体材料的材料，诸如来自例如粪便样品、组织样品或组织活检的材料。术语“生物样品”还包括从组织培养或细胞培养得到的材料。可采用任何用于获得生物样品的合适方法;示例性方法包括例如静脉切开术、拭子(例如，口腔拭子)以及细针抽吸活检程序。易于进行细针抽吸的示例性组织包括淋巴结、肺、肺洗液、BAL(支气管肺泡灌洗液)、甲状腺、乳房、胰腺和肝脏。还可以例如通过显微解剖(例如，激光捕获显微解剖(LCM)或激光显微解剖(LMD))、膀胱冲洗、涂片(例如，PAP涂片)或导管灌洗来收集样品。从个体获得或得到的“生物样品”包括在从个体获得之后已经以任何合适的方式进行处理的任何这样的样品。

[0037] 如本文所用，“生物数据”是指从生物样品得到的任何数据。这样的生物数据包括但不限于利用特定针对蛋白质靶标的适配体、任选地在基于适配体的多重测定中收集的蛋白质组学数据。

[0038] 如本文所用，“临床因素”是指可能与疾病病状或事件的风险增大相关联的生理属

性。临床因素包括但不限于年龄、体重、血压、身高、BMI、胆固醇和性别。

[0039] 如本文所用，“类别不平衡”是指数据集的一个特性，它描述了当所述集的数据被分类为两个或更多个类别时，这两个或更多个类别具有基本不相等数量的观察结果。

[0040] 如本文所用，“交叉验证”是指用于评估用来建立模型的数据上的模型性能以及统计分析的结果将如何推广到独立数据集的任何模型建立和验证技术，包括但不限于k折交叉验证、蒙特卡洛交叉验证和排除p验证(其中p可以是1至样本总数-1)。

[0041] 如本文所用，“降采样”是指对具有更多观察结果的类别(即，多数数据类别)的数据取子集，以减少类别不平衡。

[0042] 如本文所用，“相等”或“基本相等”是指在观察结果数量上具有小于10％的差异的比较类别之间的差异。

[0043] 如本文所用，“特征”是指数据集中受试者的可测量性质或特性。特征包括但不限于蛋白测量结果和临床因素。

[0044] 如本文所用，“多数数据类别”是指具有两个类别的类别不平衡数据集中具有数量更多的观察结果的那个类别。

[0045] 如本文所用，“少数数据类别”是指具有两个类别的类别不平衡数据集中具有数量更少的观察结果的那个类别。

[0046] 如本文所用，“生存分析”是指事件发生时间数据的任何建模。生存分析方法可用于任何事件发生时间(time-to-event)结果，例如MI发生时间、糖尿病发病、各种形式的癌症的发病等。生存分析包括但不限于Cox比例风险分析、随机森林分析和加速失效时间分析。

[0047] 如本文所用，“生存数据集”是指包括时间值和事件状态值两者的任何数据集，所述时间值和事件状态值指示所关注事件是否在观察受试者的时间段内发生。

[0048] 在生存分析中，类别不平衡造成的主要问题是：在一定的时间范围内，没有疾病(或事件)的个体数量超过患有疾病(或事件)的个体数量。这种不平衡可能会导致对疾病风险更高的个体的风险预测不准确。降采样通过平衡少数类别和多数类别中的个体数量来缓解这个问题，从而改进与少数类别中的个体相关的特征的检测和选择以及它们对疾病或事件发生风险的估计影响。

[0049] 已经证明对用于生存分析的类别不平衡数据集进行降采样可以改进AUC的一种情况是使用由SOMAscan®蛋白质组学测定生成的蛋白质组学数据，所述蛋白质组学数据用于识别与稳定型冠心病(CHD)患者的心血管事件风险相关联的循环蛋白生物标志物。所得模型提供了相比现有临床风险工具有所改进的能力，并且在心血管事件的复合终点中具有广泛的适用性和可推广性。

[0050] 本公开描述了用于预测稳定型CHD患者中的继发性MI的针对性模型。蛋白质组学数据用于识别稳定型CHD患者中的在抽血后四年内可能经历继发性MI的患者。除了蛋白质组学信号外，数据还包含有关在观察过程中是否发生了特定心血管事件的信息，以及发生以下情况的时间的长度：a)事件或b)由于其他因素而退出研究。这些事件发生时间数据使问题非常适合于生存分析技术。

[0051] 当主要目标是正确识别将在4年内发生MI事件的个体时，可将分析重新构造为一种分类问题，其中如果事件在4年之前发生，则个体为“阳性”类别，而如果个体留在研究中

超过4年的时间范围而没有MI，则将个体标记为“阴性”类别。生存分析工具的使用改进了模型的预测准确性(与标准分类模型相比)，因为生存模型通过将MI发生时间合并到分类器的开发中来“使用所有信息”。这种重新构造还允许使用标准分类指标(诸如AUC和混淆矩阵)来评估模型性能。这种评估生存模型的方法不是传统方法，但是特定于事件的分类在临床环境中提供了许多好处。在广泛的受众中更容易理解将患者标记为“阳性”或“阴性”(例如，与风险比或概率相比)。对预后测试的这种更好的理解使临床医生可以提供更精确、更有针对性的医疗管理。但是，与标准分类建模一样，这种生存分析方法可能会遭受经历或未经历事件的患者的不平衡。

[0052] 例如，在实施例1中所分析的亚群中，只有8.1％的个体在4年内发生继发性MI，但数量超过八倍的参与者(66.9％)无事件生存超过四年。降采样的目的是“偏置”分类器，以便它对已诊断和未诊断的个体给予同等关注，以平衡模型的敏感性和特异性。重采样技术已应用于各种机器学习方法中，但是，在使用生存建模技术的机器学习中，类别不平衡是一个尚未探讨的主题。

[0053] 在实施例1中，将降采样与Cox比例风险弹性网回归模型相结合，并且评估了对4初次抽血后4年内发生MI事件的预测。

[0054] 从实施例1明显看出，通过在建模过程中对数据进行降采样，可改进生存分析(例如Cox比例风险弹性网模型(即“Coxnet”模型))的性能。本公开有效地证明降采样后的Coxnet模型优于标准Coxnet模型、降采样后的弹性网逻辑回归模型和标准弹性网逻辑回归模型。

[0055] 除了降采样外，还有其他用于处理类别不平衡的方法也可以合并到生存模型中。例如，可结合传统的生存分析以及扩展杂的机器学习方法(诸如随机生存森林)来考虑案例加权、简单的过采样或更复杂的过采样技术(诸如合成少数类过采样技术(SMOTE))。

[0056] 尽管实施例1详细描述了在预测指定时间范围内的MI事件的情况下在生存分析中结合降采样，但是本文公开的方法可应用于对选定时间范围内的疾病病状或疾病相关事件风险的任何预测。

[0057] 图1是根据本公开的各方面的网络化计算环境100的框图，所述网络化计算环境100用于例如通过对类别不平衡数据进行降采样来处理电子数据以确定疾病风险。如图1所示，网络化计算环境100可包括疾病风险分析平台102，所述疾病风险分析平台102包括服务器系统104和电子数据库106。服务器系统104可存储并执行用于通过诸如因特网的电子网络108来使用的疾病风险分析平台102的软件模块、算法或其他子系统。用户可通过诸如计算装置等用户装置110来通过电子网络108访问疾病风险分析平台102。用户装置110可允许用户显示网络浏览器以用于通过电子网络108访问由服务器系统104托管的疾病风险分析平台102。用户装置110可以是用于访问网页的任何类型的装置，诸如个人计算装置、移动计算装置等。源装置112可通过电子网络108向疾病风险分析平台102提供数据和/或从疾病风险分析平台102接收数据。源装置112可以是用于访问网页的任何类型的装置，诸如个人计算装置、移动计算装置等。

[0058] 仅作为示例提供了图1。其他示例是可能的，并且可能不同于图1的网络化计算环境100。另外，作为示例提供了在网络化计算环境100中示出的装置和网络的数量和布置。实际上，网络化计算环境100中可能有附加的装置、更少的装置和/或网络、不同的装置和/或

网络、或者与网络化计算环境100中示出的那些布置不同的装置和/或网络。此外，图1中示出的两个或更多个装置可在单个装置内实现，或者图1中示出的单个装置可实现为多个分布式装置。另外或替代地，网络化计算环境100的一个或多个用户装置和/或服务器系统可执行服务器系统104和/或疾病风险分析平台102的一个或多个功能。

[0059] 图2描绘了用于处理电子数据以确定疾病风险的示例性计算机架构200。具体地，图2描绘了根据本公开的一个或多个实施方案的示例性计算机架构200，所述计算机架构200被配置用于将类别不平衡集的降采样与生存分析相结合。如图2的计算机架构200中所示，疾病风险分析平台102的服务器系统104可包括数据获取模块212、降采样模块214和交叉验证模块216。疾病风险分析平台102还可包括一个或多个本地或远程访问的数据库或数据存储。例如，如图2所示，疾病风险分析平台102可包括类别不平衡数据集206，所述类别不平衡数据集206包括多数类别数据202和少数类别数据204。疾病风险分析平台102还可包括降采样后的数据集208和生存模型210。应当理解，数据获取模块212、降采样模块214、交叉验证模块216、类别不平衡数据集206、降采样后的数据集208和生存模型210中的一个或多个可使其功能和内容中的一些或全部在本地、远程或同时在本地和远程进行存储或执行，并且其功能可跨平台的其他部件进行组合或分布。

[0060] 在示例性计算机架构200的一个实施方案中，数据获取模块212可从用户装置110或源装置112接收包括多数类别数据202和少数类别数据204的类别不平衡数据集206。这个类别不平衡数据集206可由降采样模块214处理以产生降采样后的数据集208。降采样后的数据集208可由交叉验证模块216处理以产生生存模型210。这个生存模型210然后可通过电子网络108发送到用户装置100和/或源装置112。

[0061] 如果使用可编程逻辑，则这样的逻辑可在可商购获得的处理平台或专用装置上执行。本领域的普通技术人员可理解，可利用各种计算机系统配置来实践所公开的主题的实施方案，这些计算机系统配置包括多核多处理器系统、小型计算机、大型计算机、与分布式功能链接或集群的计算机，以及几乎可嵌入到任何装置中的普适计算机或微型计算机。

[0062] 例如，至少一个处理器装置和存储器可用于实现上述实施方案。处理器装置可以是单个处理器、多个处理器或其组合。处理器装置可具有一个或多个处理器“核心”。

[0063] 如上文在图1和图2的示例中所描述的的本公开的各种实施方案可使用处理器装置来实现。在阅读此描述之后，对相关领域的技术人员来说变得显而易见的是，如何使用其他计算机系统和/或计算机架构来实现本公开的实施方案。尽管可将操作描述为顺序过程，但这些操作中的一些实际上可并行地、同时和/或在分布式环境中执行，并且程序代码在本地或远程存储以供单处理器或多处理器机器访问。另外，在一些实施方案中，可重新安排操作的顺序而不背离所公开的主题的精神。

[0064] 应当理解，疾病风险分析平台102和/或用于访问疾病风险分析平台102的任何装置(诸如用户装置110或源装置112)可包括中央处理单元(CPU)。这样的CPU可以是任何类型的处理器装置，包括例如任何类型的专用或通用微处理器装置。如相关领域的技术人员将理解的，CPU也可以是多核/多处理器系统中的单个处理器，这样的系统单独操作或在计算装置的群集中操作，这些计算装置在集群或服务器场中操作。CPU可连接到数据通信基础结构，例如总线、消息队列、网络或多核消息传递方案。

[0065] 还应当理解，疾病风险分析平台102和/或用于访问疾病风险分析平台102的任何

装置(诸如用户装置110或源装置112)还可包括主存储器,例如随机存取存储器(RAM),并且还可包括辅助存储器。辅助存储器,例如只读存储器(ROM),可以是例如硬盘驱动器或可移动存储驱动器。这样的可移动存储驱动器可包括例如软盘驱动器、磁带驱动器、光盘驱动器、闪速存储器等。本示例中的可移动存储驱动器以众所周知的方式读取和/或写入可移动存储单元。可移动存储单元可包括由可移动存储驱动器读取和写入的软盘、磁带、光盘等。如相关领域的技术人员将理解的,这样的可移动存储单元通常包括其中存储有计算机软件和/或数据的计算机可用存储介质。

[0066] 在替代实现方式中,辅助存储器可包括用于允许将计算机程序或其他指令加载到装置中的其他类似装置。这样的装置的示例可包括:程序盒和盒式接口(诸如在视频游戏装置中所发现的)、可移动存储器芯片(诸如EPROM或PROM)和相关联的插槽,以及允许软件和数据从可移动存储单元传输到装置的其他可移动存储单元和接口。

[0067] 还应当理解,疾病风险分析平台102和/或用于访问疾病风险分析平台102的任何装置(诸如用户装置110或源装置112)还可包括通信接口(“COM”)。通信接口允许在装置和外部装置之间传输软件和数据。通信接口可包括调制解调器、网络接口(诸如以太网卡)、通信端口、PCMCIA插槽和卡等。通过通信接口传输的软件和数据可以呈信号的格式,所述信号可以是电子信号、电磁信号、光信号或能够由通信接口接收的其他信号。这些信号可通过装置的通信路径被提供给通信接口,所述通信路径可使用例如电线或电缆、光纤、电话线、蜂窝电话链路、RF链路或其他通信信道来实现。

[0068] 这样的设备的硬件元件、操作系统和编程语言本质上是常规的,并且假设本领域技术人员对此足够熟悉。用于访问疾病风险分析平台的装置还可包括输入和输出端口,以与输入和输出装置(诸如键盘、鼠标、触摸屏、监视器、显示器等)连接。当然,各种服务器功能可在许多类似平台上以分布式方式实现,以便分布处理负载。替代地,服务器可通过一个计算机硬件平台的适当编程来实现。

[0069] 通过示例并参考附图来详细描述本文公开的系统、设备、装置和方法。本文讨论的示例仅是示例,并且被提供以帮助解释本文描述的设备、装置、系统和方法。对于所述设备、装置、系统或方法中的任何一个的任何特定实现方式,除非特别指定为强制性的,否则附图中示出或以下讨论的特征或部件均不应被视为强制性的。为了易于阅读和清楚起见,可仅结合特定附图来描述某些部件、模块或方法。在本公开中,对特定技术、布置等的任何标识与所呈现的特定示例有关,或者仅仅是对这种技术、布置等的一般描述。除非明确这样指定,否则对特定细节或示例的标识不意图且不应当被理解为强制性或限制性的。未能具体描述部件的组合或子组合的任何失败不应被理解为指示任何组合或子组合是不可能的。将理解的是,可对所公开和描述的示例、布置、配置、部件、元件、设备、装置、系统、方法等进行修改,并且这些修改对于特定应用可能是所希望的。同样,对于所描述的任何方法,无论是否结合流程图描述了所述方法,都应理解,除非上下文另外指定或要求,否则在执行方法时进行的步骤的任何显式或隐式排序都不意味着这些步骤必须按所呈现的顺序进行,而是可按不同的顺序或并行地进行。

[0070] 在本公开全篇中,对部件或模块的引用通常是指可在逻辑上被分组在一起以进行一个功能或一组相关功能的项目。部件和模块可以用软件、硬件或软件和硬件的组合来实现。术语“软件”被广泛地使用以不仅包括可执行代码,例如机器可执行或机器可解释的指

令,而且包括以任何合适的电子格式存储的数据结构、数据存储和计算指令,包括固件和嵌入式软件。术语“信息”和“数据”被广泛地使用,并且包括:各种各样的电子信息,包括可执行代码;诸如文本、视频数据和音频数据之类的内容;以及各种代码或标志。当上下文允许时,术语“信息”、“数据”和“内容”有时可互换使用。

[0071] 实施例

[0072] 提出以下实施例以便更全面地说明本发明的一些实施方案。然而,它们决不应被理解为限制本发明的广泛范围。本领域的普通技术人员可容易地采用此发现的基本原理来设计多种化合物而不背离本发明的精神。

[0073] 实施例1

[0074] 本实施例提供的描述是将降采样与Cox比例风险弹性网回归模型相结合,以评估对初次抽血后4年内发生心肌梗死(MI)事件的预测,这可以在图2的示例性数据风险分析平台内完成。

[0075] 本实施例的目的至少有两方面:1)选择和识别可预测少数类别和多数类别两者的特征,以及2)推导估计效应大小,使得可很好地预测少数类别的风险。作为对比,检验了逻辑回归弹性网模型(有和没有降采样)的预测能力以及没有降采样的Cox弹性网模型的预测能力。

[0076] 材料和方法-数据集

[0077] 分析中使用的样品是来自HUNT3研究的亚群,HUNT3研究是来自挪威的一项前瞻性人群研究,它包括从研究参与者身上抽取的血液样品和后续的健康信息。CHD亚群先前已有描述(Peter Ganz等人的Development and validation of a protein-based risk score for cardiovascular outcomes among patients with stable coronary heart disease.Jama,315(23):2532-2541,2016),入选标准旨在通过超过六个月的既往、狭窄、诱导性缺血或既往冠状动脉血管重建术的MI史来证明存在但稳定的CHD。使用SOMAscan®测定(SomaLogic,Inc;Boulder,CO USA)对血浆样品进行测定,所述测定使用慢解离速率修饰适配体(SOMAmer®)试剂来测量相对蛋白质丰度。V4测定可测量5,220种蛋白质分析物,并且是完善的蛋白质生物标志物发现平台。

[0078] 在亚群中,有8.1%的患者在4年内经历了继发性MI(表1)。图3中描绘了CHD亚群中的MI的卡普兰-迈耶生存曲线。卡普兰-迈耶曲线是一种经验性的非参数方法,用于检验无事件(例如,无MI)发生的概率如何随时间变化。对于HUNT3数据集的CHD亚群中的MI,无事件发生的概率逐渐减小。表1示出CHD亚群中的MI发生率和人口统计学信息。

[0079] 表1-稳定型CHD亚群的人口统计学特性

[0080]

特性	4年内发生MI	4年内不发生MI	超过4年不发生事件(MI或其他)	总计
受试者数量(总量的%)	61 (8.1%)	189 (25.0%)	506 (66.9%)	756 (100%)
性别，女(组的%)	n=20 (32.8%)	n=57 (30.2%)	n=128 (25.3%)	n=205 (27.1%)
性别，男(组的%)	n=41 (67.2%)	n=132 (69.8%)	n=378 (74.7%)	n=551 (72.9%)
平均年龄，以岁为单位(± SD)	72.6±11.1	72.8±10.3	67.7±10.2	69.4±10.5
事件发生时间，以年为单位(IQR)	1.93 (1.97)	2.57 (2.73)	4.69 (0.75)	4.37 (1.06)

[0081] 材料和方法-Cox弹性网模型

[0082] 生存数据由结果(即,事件发生时间)来表征,它适应广泛的主题,包括MI事件、因癌症死亡、因疾病再次住院、机器部件故障等。时间相关数据的本质是,如果事件发生在研究期限之外,则对于某些个体将不会观察到这个事件。这些个体被“删失”,这可能由于多种原因而发生(例如,非MI相关原因导致的死亡、个体退出研究、MI在研究窗口结束后发生)。尽管存在多种类型的删失,但数据含有经过右删失的个体,这意味着对于没有发生MI事件的患者,假设它在最后一个观察到的时间点之后已经发生。

[0083] 生存数据通过生存函数S(.)来表征,生存函数是无事件发生的概率并且在时间点t计算为

[0084] $$S(t) = P(T > t) = \int_t^{\infty} f(u)du,$$

[0085] 其中f(.)是MI发生时间的概率密度函数。与生存函数一起,也可以识别和表征明显增大或减小事件发生时间的特征。尽管存在许多生存分析技术,但最常见的一种是Cox比例风险模型。Cox模型表示为

[0086] $\lambda(t|X_i,\beta) = \lambda_0(t)\exp\{X'_i\beta\}$.

[0087] 这里,$\lambda(t|.)$是风险函数(或“立即失效的风险”函数)并且定义为$\lambda(t|.) = f(t|.)/S(t|.)$。此外,$X_i$是第i个个体的特征测量结果的p x 1向量,而β是特征效应的p x 1向量。Cox模型的主要目标是估计特征对个体的事件发生风险的影响。基线风险率$\lambda_0(t)$在估计例程中被视为多余参数并且因此不予检验。

[0088] 由于数据集中的特征数量大于样品大小,因此可将弹性网罚分合并到我们的模型中,这是一种形式的惩罚回归,它结合了最小绝对收缩和选择算子(即,lasso)和岭回归或Tikhonov正则化。这种工具通过lasso例程进行特征选择,同时允许相关特征一起保留在模型中,使得p可大于n。在标准回归模型中,通常通过使响应Y_i和预测器$X'_i\beta$之间的差最小化来估计特征效应β。但是,通过弹性网正则化,估计的特征效应被计算为

[0089] $$\hat{\beta} = \arg\min_{\beta} |Y - X\beta|^2 + \lambda_2|\beta|^2 + \lambda_1|\beta|,$$

[0090] 其中λ_1是与lasso回归相关联的L_1罚分，并且λ_2是与岭回归相关联的L_2罚分。

[0091] 使用通过在CRAN-R中可获得的glmnet包实现的Cox弹性网模型，将生存分析与弹性网罚分相结合。Cox弹性网模型将标准Cox比例风险模型与弹性网惩罚合并在一起，从而允许使用生存技术来开发分类器，以及惩罚回归的好处。

[0092] 为了减轻类别不平衡，将Cox比例风险弹性网模型与降采样技术相结合。这种方法允许识别能够最好地预测个体是否处于在4年内发生MI事件的“高风险”的特征，其中“高风险”分类器是使用通过交叉验证识别的风险比阈值计算出的。此外，这种技术估计特征效应的方式允许准确预测高风险个体的特征所具有的“权重”(即，β估计)与使用完整人群得到的特征不同。

[0093] 为了进行比较，两个弹性网逻辑回归模型(有或没有降采样，可通过R中的插入符号包来实现)，以及没有合并降采样技术的Cox弹性网模型。适当时，使用AUC、敏感性、特异性和C指数对模型进行比较。

[0094] 在RStudio服务器版本1.1.453中使用R版本3.4.4进行分析。

[0095] 材料和方法-数据取子集

[0096] 将数据集分为训练集(数据的80％)和测试集(20％)。将训练集用于模型构建，并且在测试集上评估最终模型。用于Cox弹性网模型的测试集上的预测阈值被计算为交叉验证期间每一折所生成的阈值的平均值。在实现惩罚回归模型之前，使用训练集进行单变量过滤。计算每个分析物的学生t检验以评估在研究窗口中有和没有MI事件的个体之间的平均值在统计学上是否明显不同。在证明所述技术的效用时，为了一致，在模型开发中包括了前100种分析物(按错误发现率值排名)。

[0097] 结果

[0098] 将降采样后的Cox弹性网模型的结果与两个逻辑回归弹性网模型(降采样后的和未降采样的)以及没有使用降采样的Cox弹性网模型进行比较。为了简化表示，将Cox弹性网模型称为“Coxnet”模型，并且将弹性网逻辑回归模型称为“LRnet”模型。对于降采样后的模型，在前面加上“DS”(例如，实现降采样的Cox弹性网模型是“DS-Coxnet”)。

[0099] 在各种模型中，对训练集使用5次重复的5折交叉验证，以在每种模型类型中选择最佳模型。通过最大AUC选择最佳模型。特征选择、估计效应和分类阈值在模型之间可以有所不同。在交叉验证后，在测试数据集上评估每个类别中的顶部模型的预测能力。

[0100] 在模型开发过程中，使用原始数据创建了Coxnet模型，但在4年的时间点处使用AUC指标对其进行优化以用于分类。这意味着建立了标准生存模型，但是使用二进制4年标记分类器(四年前有/无MI)来计算AUC并优化模型。4年结果用于开发逻辑回归模型，这些模型也使用AUC进行了优化。为了使用标准生存模型指标进行模型比较，计算了生存模型的C指数。

[0101] 模型结果和比较

[0102] 交叉验证结果显示，两种Coxnet模型都大大优于标准LRnet模型(参见表2)。这种结果是意料之中的，因为生存分析方法使用事件发生时间信息作为特征选择和模型开发的一部分。更令人信服的结果是，在所有分类指标(AUC、敏感性、特异性)上，DS-Coxnet模型均优于DS-LRnet模型和标准Coxnet模型两者。此外，DS-Coxnet模型具有比标准Coxnet模型更高的C指数，这指示了降采样后的模型更好地预测MI发生时间的排序。

[0103] 表2-交叉验证后的训练集结果

[0104]

模型	调整参数	AUC	敏感性	特异性	C 指数
降采样后的Cox模型	$\alpha=0.75$，$\lambda=0.1$	0.78	0.74	0.79	0.74
Cox 模型	$\alpha=0.75$，$\lambda=0.05$	0.61	0.67	0.66	0.58
降采样后的逻辑回归	$\alpha=0.75$，$\lambda=0.05$	0.75	0.65	0.73	-
逻辑回归	$\alpha=0.75$，$\lambda=0.05$	0.58	0	1	-

[0105] 在通过交叉验证进行模型优化后，在测试集上评估顶部模型的预测能力，包括基于通过4年标记将个体正确地预测为有发生MI的“高风险”来检验敏感性和特异性。表3中示出测试集上所有模型的性能指标。DS-Coxnet模型是唯一表现优于“随机机会”且AUC为0.63的模型。此外，与DS-LRnet模型和标准Coxnet模型两者相比，DS-Coxnet模型具有最高的敏感性和特异性(不出所料，LRnet模型在测试数据集上的表现与在训练数据集上的表现一样差)。

[0106] 表3-测试集结果

[0107]

模型	阈值	AUC	敏感性	特异性	C指数
降采样后的Cox模型	0.46	0.63	0.46	0.80	0.74
Cox模型	-0.004	0.49	0.38	0.56	0.49
降采样后的逻辑回归	0.50	0.54	0.15	0.72	-
逻辑回归	0.50	0.49	0	1	-

[0108] 为了进一步证明降采样后的生存模型方法的好处，对于每个模型，在测试集上生成卡普兰-迈耶曲线，使用通过交叉验证识别的特定于模型的阈值来预测个体是否为高风险会使所述曲线分层(参见图4)。为了进行这种比较，将标准Coxnet模型和DS-Coxnet模型的阈值计算为交叉验证迭代中的平均阈值。这种视觉检查方法使用DS-Coxnet模型的阈值显示了高风险组和平均风险组之间的非常明确的分离。对于其他模型，这种分离的定义不是很明确。

[0109] 数字和模型评估指标的组合证据(表3)提出了一个令人信服的案例，即降采样后的生存模型方法有益于识别处于四年内发生MI的高风险的个体。

[0110] 降采样后的Coxnet模型的阈值研究

[0111] 使用DS-Coxnet模型对测试集进行预测所使用的阈值是交叉验证迭代中的所有阈值的平均值。尽管此阈值导致比其他模型更高的敏感性和特异性，但这些值仍然相当不平衡。一个重要的考虑因素是敏感性/特异性的权衡是否可通过操纵预测阈值进一步得到平衡。

[0112] 与分类模型一样，阈值可被调整以找到在测试集上使敏感性最大、使特异性最大或使敏感性和特异性之间的差异最小的值。表4显示了测试集上不同阈值的性能指标，并且图5绘制了每种情况的卡普兰-迈耶曲线。如表4中所示，改变预测阈值会导致敏感性高于60％，而AUC不会减小。但是，在使用平均阈值的情况下，卡普兰-迈耶曲线(图5)显示了高风险个体和平均风险个体之间的最宽分离。

[0113] 表4-使用降采样后的Cox模型在测试集上应用不同的阈值

[0114]

实验	阈值	AUC	敏感性	特异性
平均阈值	0.46	0.63	0.46	0.80
平衡的敏感性与特异性	0.165	0.63	0.62	0.64
使敏感性最大	0.165	0.63	0.62	0.64
使特异性最大	0.712	0.58	0.31	0.85

[0115] 尽管敏感性和特异性仍相对低于通常所需的水平(即,70%或更高),但此结果可能是由于以下事实:测试集中只有13位受试者在四年前发生了MI事件,这限制了模型的开发。但是,分析证明了可采用与分类模型相同的方式来调整用于对生存模型中的风险水平进行分类的阈值。

[0116] 意图仅将说明书和实施例视为示例性的,而本公开的真实范围和精神由以下权利要求书指示。

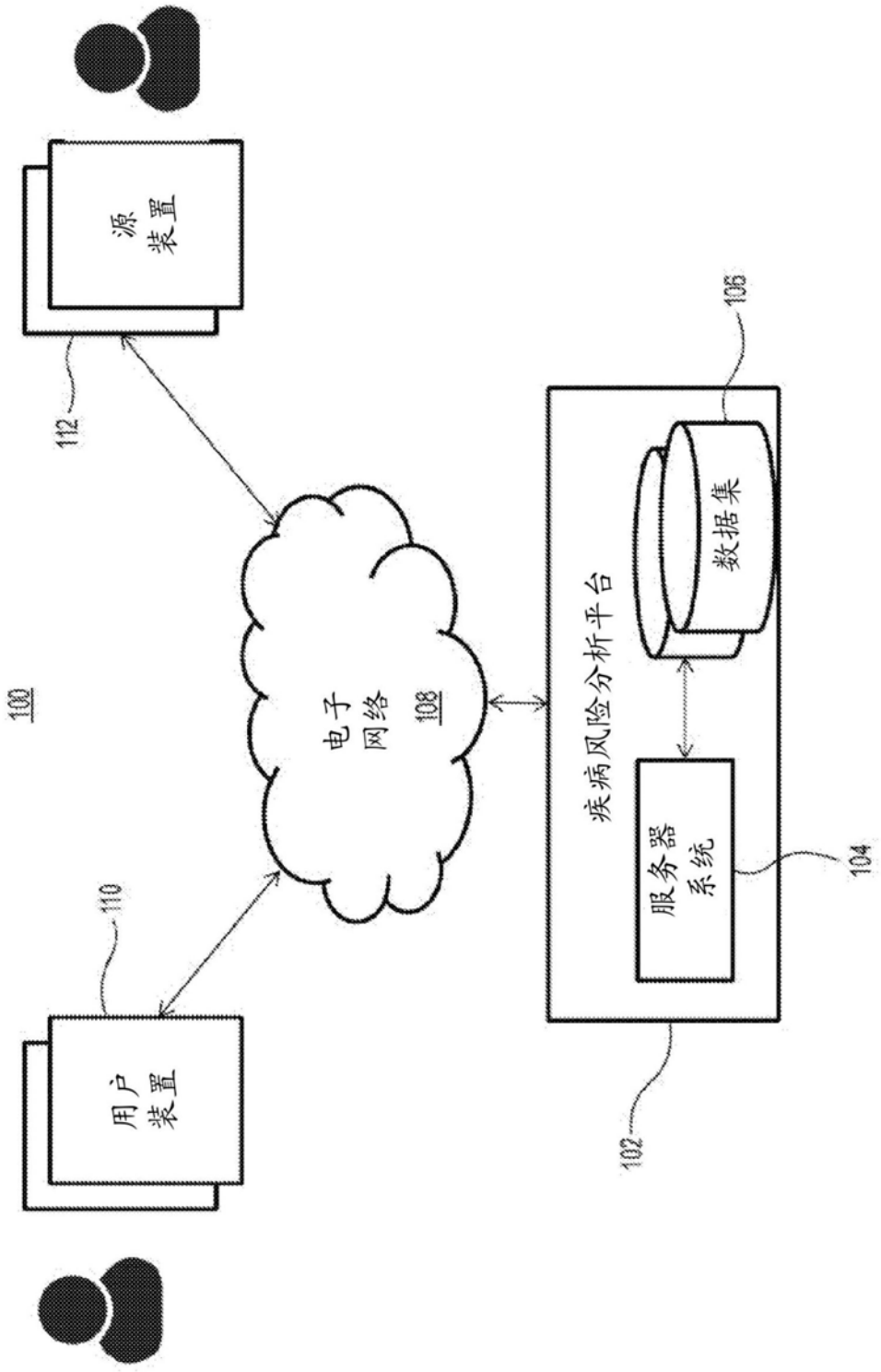


图1

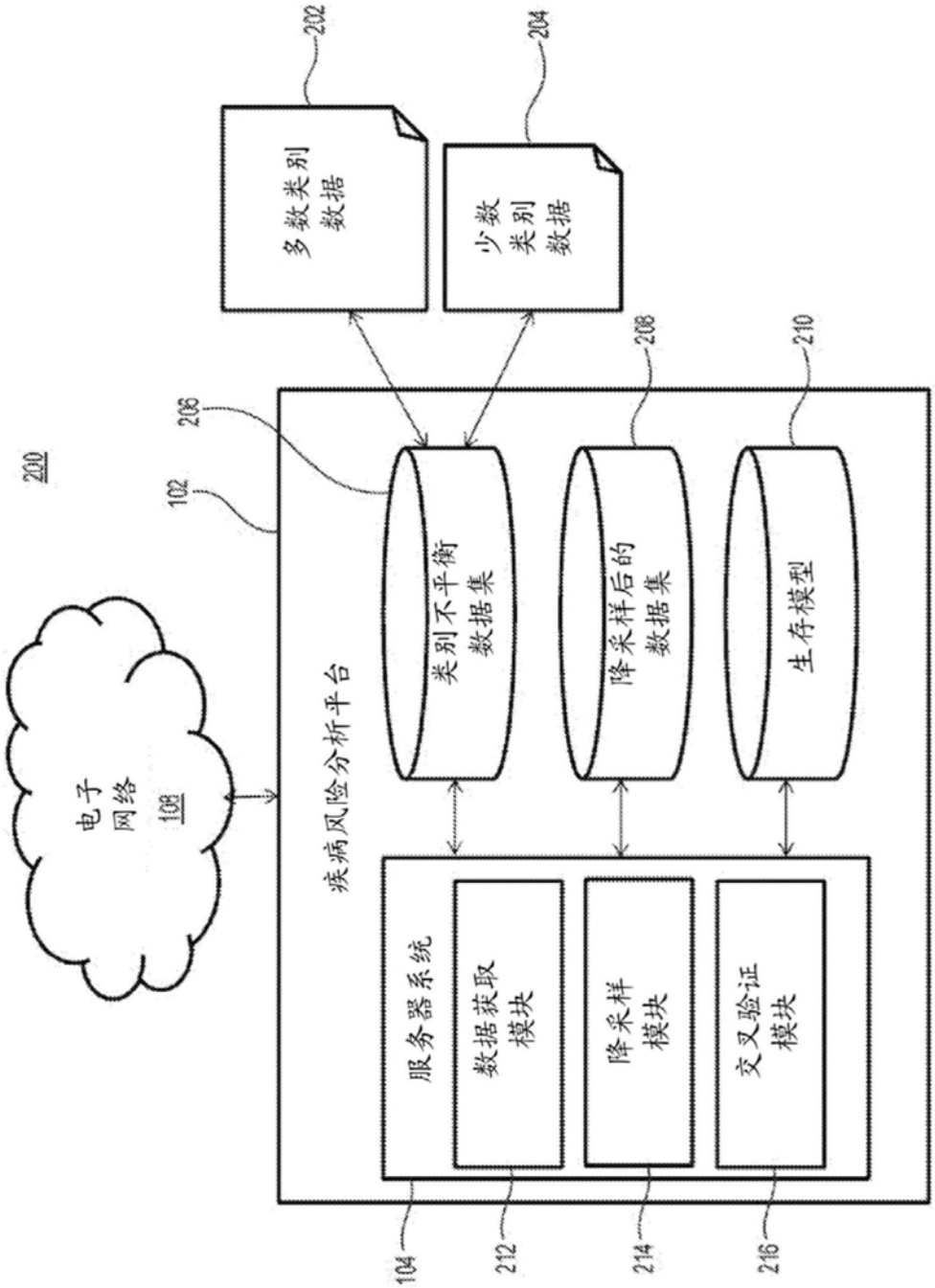

200
202
多数类别数据
204
少数类别数据
206
102
208
210
类别不平衡数据集
降采样后的数据集
生存模型
电子网络 108
疾病风险分析平台
服务器系统
数据获取模块
降采样模块
交叉验证模块
104
212
214
216

图2

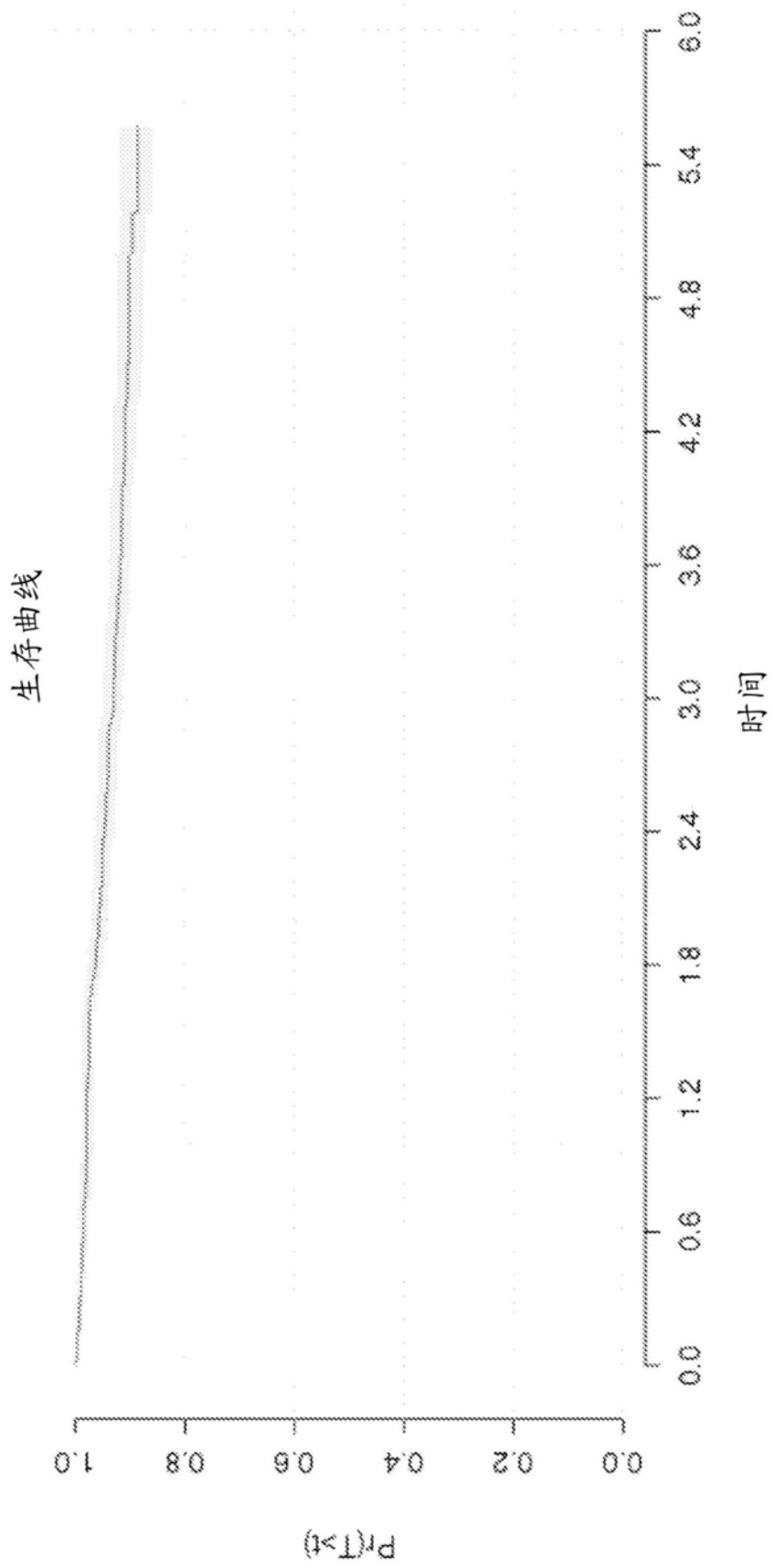


图3

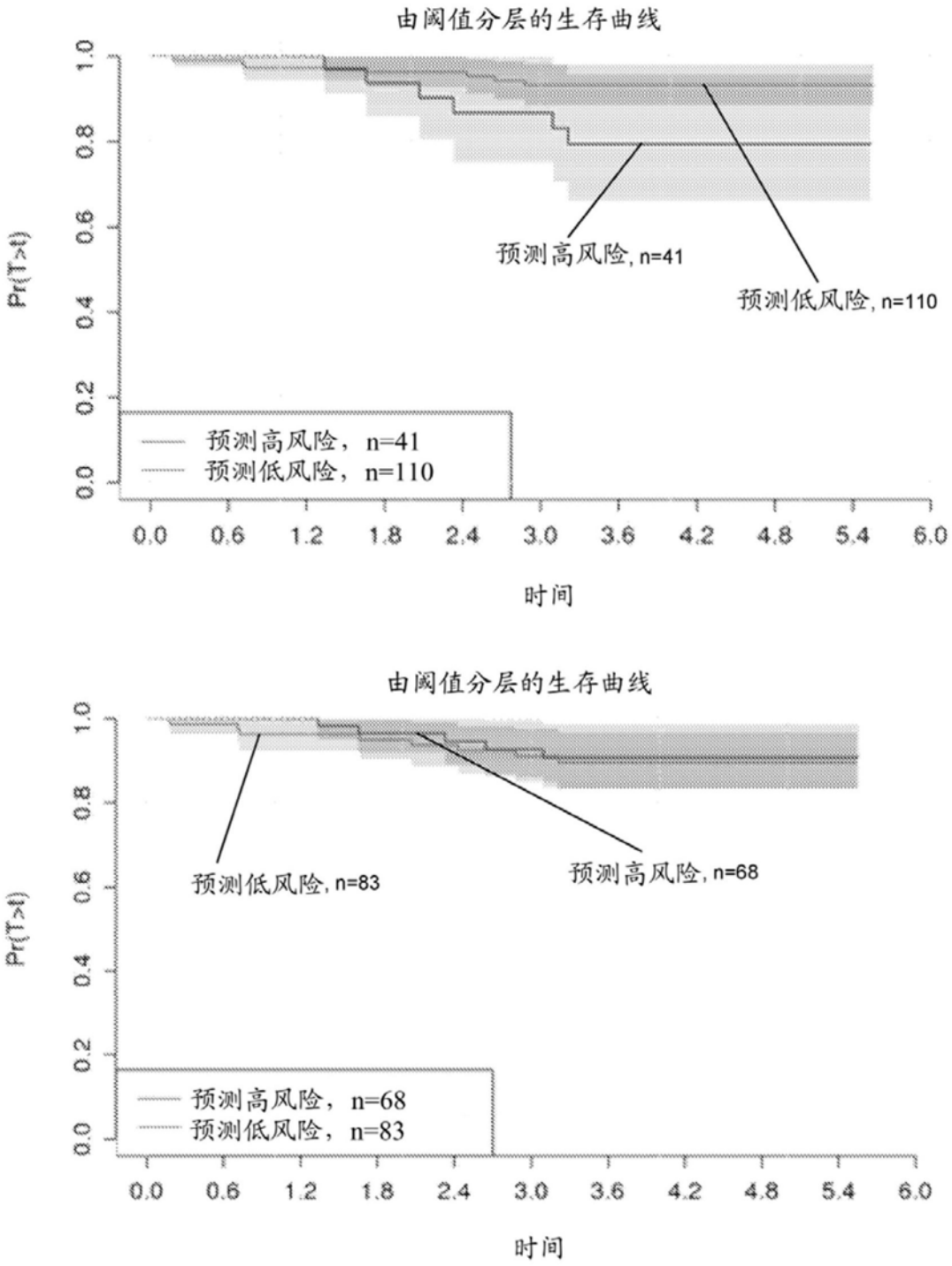
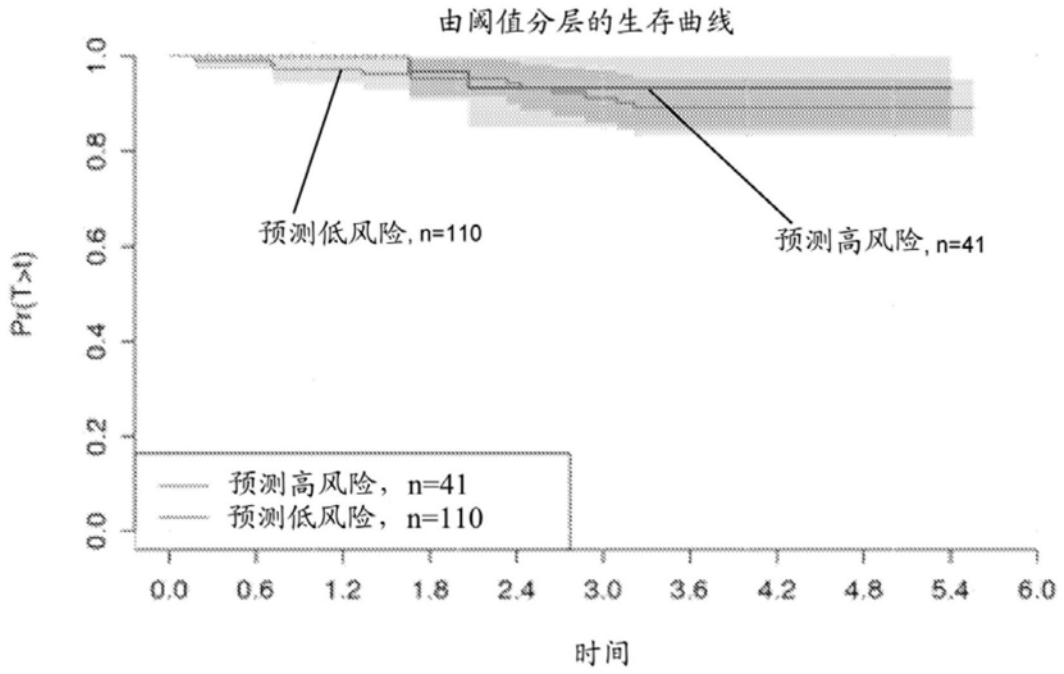

由阈值分层的生存曲线
Pr(T>t)
1.0
0.8
0.6
0.4
0.2
0.0
预测高风险, n=41
预测低风险, n=110
预测高风险， n=41
预测低风险， n=110
0.0
0.6
1.2
1.8
2.4
3.0
3.6
4.2
4.8
5.4
6.0
时间
由阈值分层的生存曲线
Pr(T>t)
预测低风险, n=83
预测高风险, n=68
预测高风险， n=68
预测低风险， n=83
时间

图4


由阈值分层的生存曲线
预测低风险, n=110
预测高风险, n=41
Pr(T>t)
1.0
0.8
0.6
0.4
0.2
0.0
预测高风险， n=41
预测低风险， n=110
0.0 0.6 1.2 1.8 2.4 3.0 3.6 4.2 4.8 5.4 6.0
时间

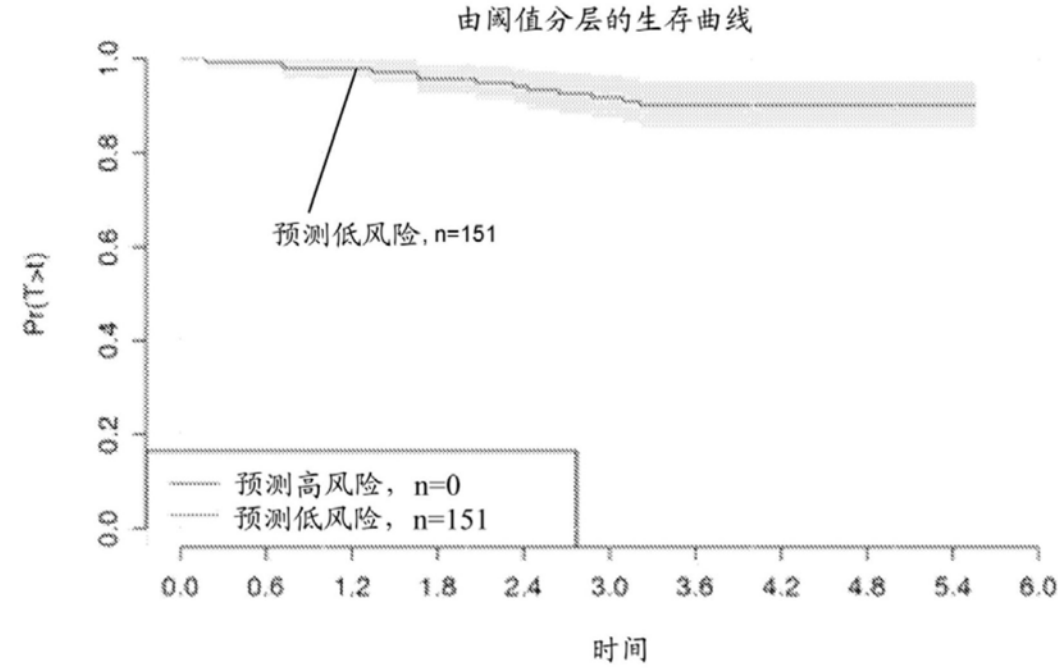
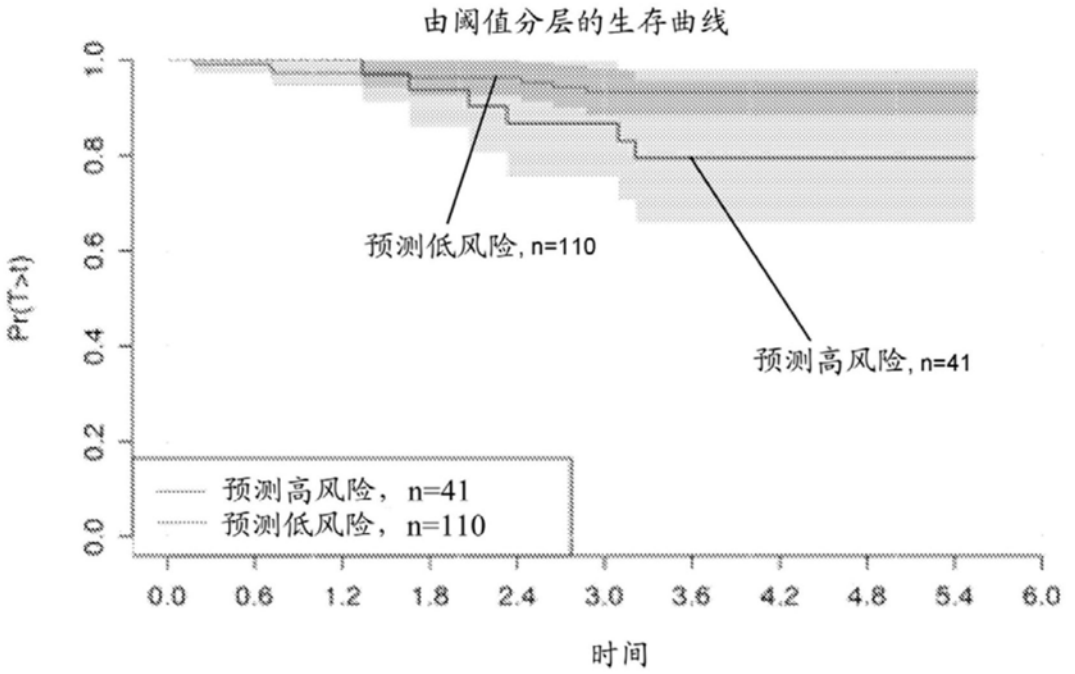

由阈值分层的生存曲线
预测低风险, n=151
Pr(T>t)
1.0
0.8
0.6
0.4
0.2
0.0
预测高风险， n=0
预测低风险， n=151
0.0 0.6 1.2 1.8 2.4 3.0 3.6 4.2 4.8 5.4 6.0
时间

图4续


由阈值分层的生存曲线
Pr(T>t)
1.0
0.8
0.6
0.4
0.2
0.0
预测低风险, n=110
预测高风险, n=41
预测高风险， n=41
预测低风险， n=110
0.0 0.6 1.2 1.8 2.4 3.0 3.6 4.2 4.8 5.4 6.0
时间

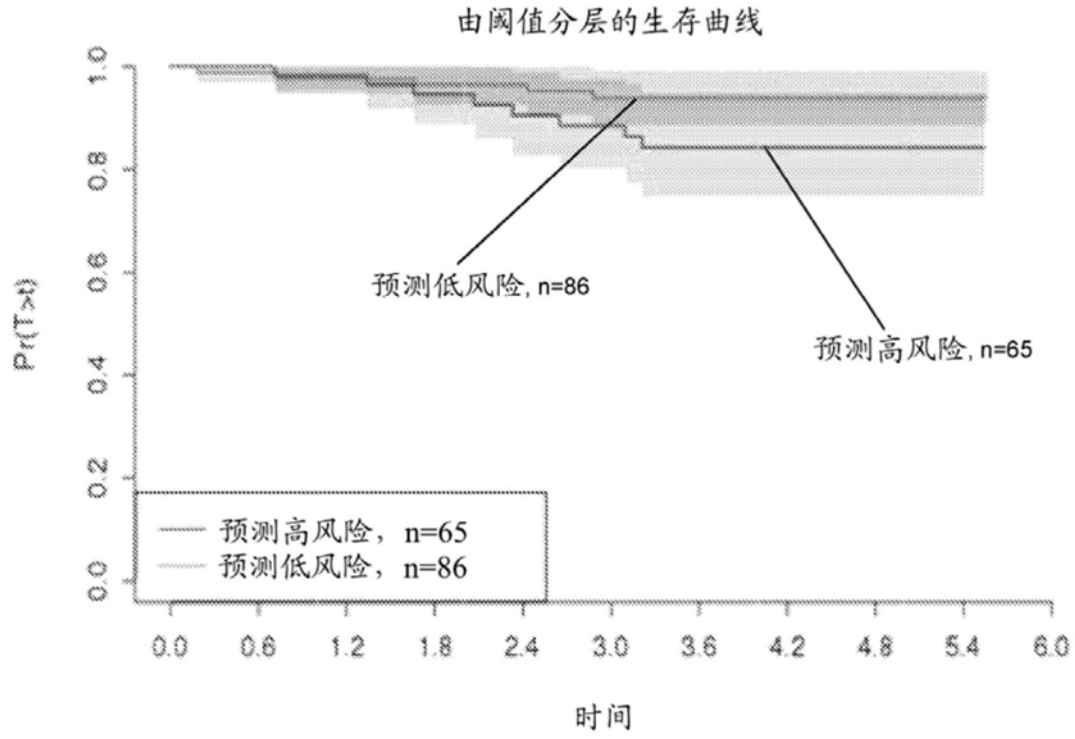

由阈值分层的生存曲线
Pr(T>t)
1.0
0.8
0.6
0.4
0.2
0.0
预测低风险, n=86
预测高风险, n=65
预测高风险， n=65
预测低风险， n=86
0.0 0.6 1.2 1.8 2.4 3.0 3.6 4.2 4.8 5.4 6.0
时间

图5

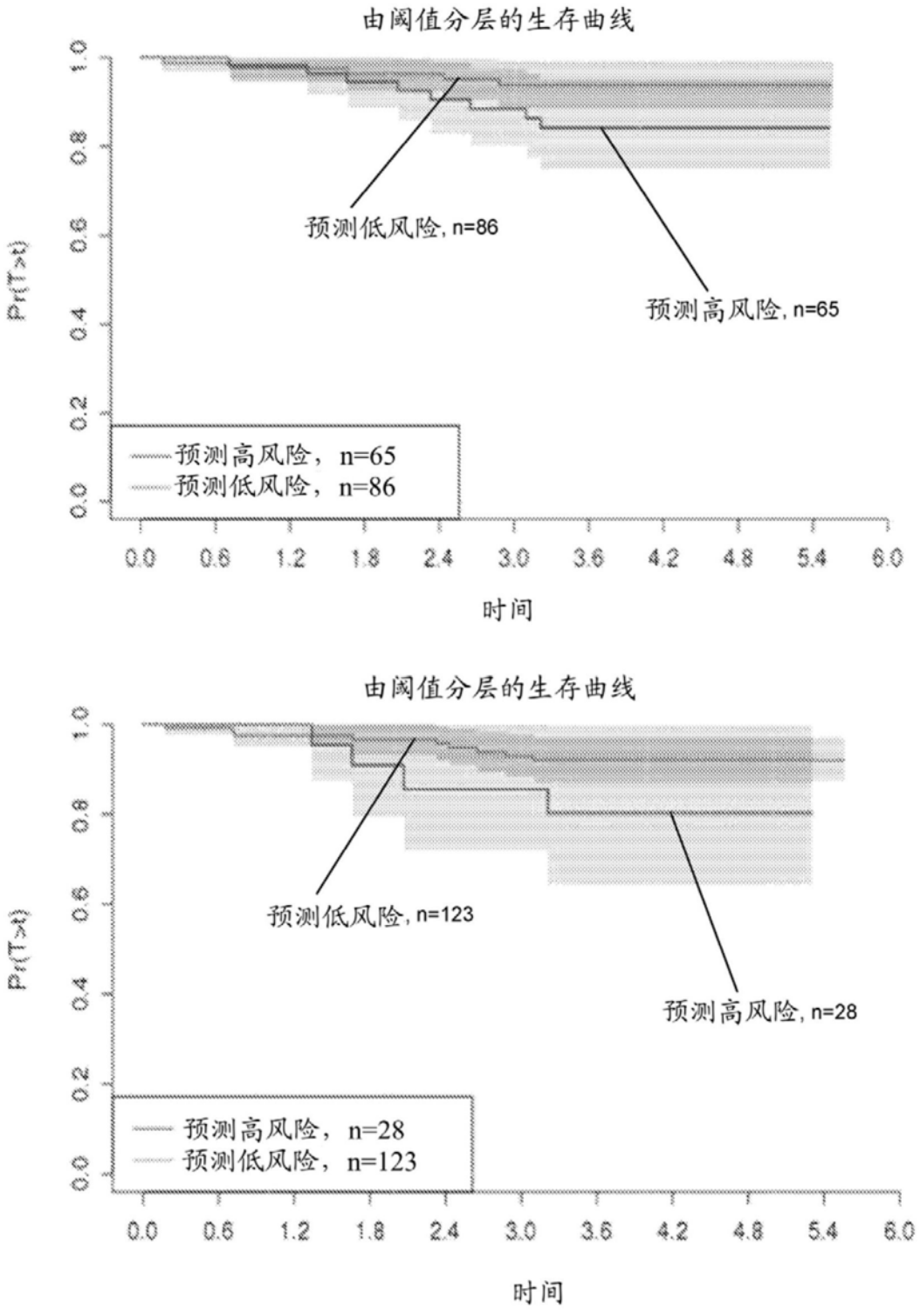

由阈值分层的生存曲线
Pr(T>t)
1.0
0.8
0.6
0.4
0.2
0.0
预测低风险, n=86
预测高风险, n=65
预测高风险， n=65
预测低风险， n=86
0.0
0.6
1.2
1.8
2.4
3.0
3.6
4.2
4.8
5.4
6.0
时间
由阈值分层的生存曲线
Pr(T>t)
预测低风险, n=123
预测高风险, n=28
预测高风险， n=28
预测低风险， n=123
时间

图5续