(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 104756117 B
(45)授权公告日 2019.01.29

(21)申请号 201380055556.4

(22)申请日 2013.10.17

(65)同一申请的已公布的文献号
申请公布号 CN 104756117 A

(43)申请公布日 2015.07.01

(30)优先权数据
61/718,242 2012.10.25 US

(85)PCT国际申请进入国家阶段日
2015.04.23

(86)PCT国际申请的申请数据
PCT/IB2013/059424 2013.10.17

(87)PCT国际申请的公布数据
WO2014/064585 EN 2014.05.01

(73)专利权人 皇家飞利浦有限公司
地址 荷兰艾恩德霍芬

(72)发明人 B·J·巴克 H·J·范奥义任 R·范德汉姆

(74)专利代理机构 永新专利商标代理有限公司 72002
代理人 李光颖 王英

(51)Int.Cl.
G16H 10/20(2018.01)

(56)对比文件
US 2009/0089079 A1,2009.04.02,
高丹丽 等.剖宫产术后下肢深静脉血栓形成的危险因素分析.《中国妇幼保健》.2012,第27卷(第23期),第3581-3584页.

审查员 李宝

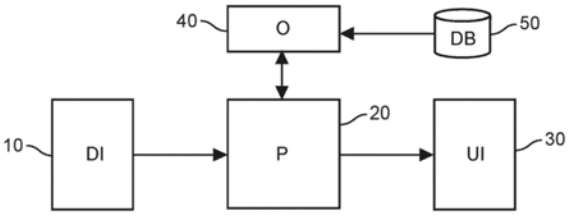
权利要求书3页 说明书11页 附图5页

(54)发明名称

用于临床决策支持的对血栓形成的临床风险因子与分子标记物的组合使用

(57)摘要

本发明涉及一种用于临床决策支持的装置和方法,所述装置和方法基于临床风险因子和诸如蛋白质浓度的分子标记物的组合来识别处于高血栓形成风险中的患者。这些临床风险因子和分子标记物被组合在基于机器学习的算法中,所述基于机器学习的算法返回与未来血栓形成事件的估计的风险相关的输出值。



CN 104756117 B

1.一种用于基于患者特异性输入特征来计算患者的血栓形成风险的估计值的装置,所述装置包括:

数据接口(10、30),其用于接收所述输入特征;

处理器,其用于通过应用根据从所述接收的输入特征导出的数值的决策支持算法来计算所述估计值;以及

用户接口(30),其用于输出所述估计值;

其中,所述输入特征包括所述患者的至少一个临床风险因子和至少一个蛋白质浓度的组合,并且

其中,所述至少一个临床风险因子是从以下中选择的:

在第一预定时间段内的固定、

在第二预定时间段内的外科手术、

静脉血栓形成的家族史、

在第三预定时间段内的妊娠或分娩、

对雌激素的当前使用,以及

肥胖。

2.根据权利要求1所述的装置,其中,所述至少一个蛋白质浓度是从以下中选择的:血液中的凝血蛋白FVIII的浓度、血液中的凝血蛋白FXI的浓度、以及血液中的凝血蛋白TFPI的浓度。

3.根据权利要求1所述的装置,其中,所述第一预定时间段对应于至少三个月,所述第二预定时间段对应于一个月,并且所述第三预定时间段对应于至少三个月。

4.根据权利要求1所述的装置,其中,所述处理器(20)适于将所述估计值与预定阈值相比较并且基于比较结果来对所述估计值进行分类。

5.根据权利要求4所述的装置,其中,所述装置适于允许用户输入所述预定阈值或禁用所述预定阈值。

6.根据权利要求1所述的装置,还包括优化单元(40),所述优化单元用于基于在数据库(50)中存储的数据集通过优化流程来应用学习过程,从而使预测误差最小化。

7.根据权利要求1所述的装置,其中,所述处理器(20)适于基于临床风险因子、单核苷酸多态性以及蛋白质水平来计算深静脉血栓形成风险得分。

8.一种用于基于患者特异性输入特征来计算患者的血栓形成风险的估计值的装置,所述装置包括:

用于选择所述输入特征以包括所述患者的至少一个临床风险因子和至少一个蛋白质浓度的组合的单元;

用于通过应用根据从所述输入特征导出的数值的决策支持算法来计算所述估计值的单元;以及

用于从以下中选择所述至少一个临床风险因子的单元:

在第一预定时间段内的固定、

在第二预定时间段内的外科手术、

静脉血栓形成的家族史、

在第三预定时间段内的妊娠或分娩、

对雌激素的当前使用,以及

肥胖。

9.根据权利要求8所述的装置,还包括用于通过基于多个患者的存储的数据集通过学习过程来优化所述输入特征从而使预测误差最小化的单元。

10.根据权利要求9所述的装置,还包括:用于将所述数据集划分为训练集、验证集和测试集的单元;用于使用所述训练集和所述验证集来选择机器学习函数的类型和用于优化分类器的模型参数集的单元;用于使用经优化的分类器来获得所述患者特异性输入特征的单元;以及用于使用所述测试集基于所述获得的输入特征来计算针对所述测试集的患者的所述估计值的单元。

11.根据权利要求8所述的装置,还包括用于从以下中选择所述至少一个蛋白质浓度的单元:血液中的凝血蛋白FVIII的浓度、血液中的凝血蛋白FXI的浓度、以及血液中的凝血蛋白TFPI的浓度。

12.根据权利要求8所述的装置,还包括用于将所述第一预定时间段设置为至少三个月、将所述第二预定时间段设置为一个月、并且将所述第三预定时间段设置为至少三个月的单元。

13.一种计算机可读介质,其上存储了程序代码模块,当所述程序代码模块在计算机设备上运行时,所述程序代码模块用于令计算机设备执行一种用于基于患者特异性输入特征来计算患者的血栓形成风险的估计值的方法的步骤,其中,所述方法包括:

选择所述输入特征以包括所述患者的至少一个临床风险因子和至少一个蛋白质浓度的组合;并且

通过应用根据从所述输入特征导出的数值的决策支持算法来计算所述估计值,

其中,所述方法还包括从以下中选择所述至少一个临床风险因子:

在第一预定时间段内的固定、

在第二预定时间段内的外科手术、

静脉血栓形成的家族史、

在第三预定时间段内的妊娠或分娩、

对雌激素的当前使用,以及

肥胖。

14.根据权利要求13所述的计算机可读介质,其中,所述方法还包括通过基于多个患者的存储的数据集通过学习过程来优化所述输入特征,从而使预测误差最小化。

15.根据权利要求14所述的计算机可读介质,其中,所述方法还包括:将所述数据集划分为训练集、验证集和测试集;使用所述训练集和所述验证集来选择机器学习函数的类型和用于优化分类器的模型参数集;使用经优化的分类器来获得所述患者特异性输入特征;并且使用所述测试集基于所述获得的输入特征来计算针对所述测试集的患者的所述估计值。

16.根据权利要求13所述的计算机可读介质,其中,所述方法还包括从以下中选择所述至少一个蛋白质浓度:血液中的凝血蛋白FVIII的浓度、血液中的凝血蛋白FXI的浓度、以及血液中的凝血蛋白TFPI的浓度。

17.根据权利要求13所述的计算机可读介质,其中,所述方法还包括:将所述第一预定

时间段设置为至少三个月，将所述第二预定时间段设置为一个月，并且将所述第三预定时间段设置为至少三个月。

用于临床决策支持的对血栓形成的临床风险因子与分子标记物的组合使用

技术领域

[0001] 本发明涉及临床决策支持领域，其中，基于患者特异性输入特征来计算患者的血栓形成风险的估计值。

背景技术

[0002] 基于计算机的临床决策支持系统(CDSS)被定义为“被设计为直接帮助临床决策制定的任何软件，在所述临床决策制定中，出于生成之后被呈现给医师以供考虑和决策制定的患者特异性评估或推荐的目的，将个体患者的特性匹配到计算机化的知识库”。临床决策支持系统通过支持临床决策制定使其改进健康护理的质量的潜力得到促进。

[0003] 深静脉血栓形成是西方世界中广泛普及的问题。大部分人口，例如老年人、旅行中的人以及经历整形外科手术的患者都处于血栓形成的增加的风险中。能够对处于风险中的人施加预防性的抗凝处置，但是出血的风险(每年1-3％)、以及成本问题以及不方便对此发言反对。因此期望具有一种更加患者特异性的措施来估计人的血栓形成风险并方便关于是否要进行处置的知情选择。遗憾的是，在当前的临床筛查技术和可用的方法的情况下，应当接受抗凝剂的高风险个体不容易被辨识并且事件没有准确地被预测。这种情况继续的主要原因之一是患有血栓形成的没有显著的遗传缺陷的绝大多数患者具有在临床上通过例行筛查工具和因子测定未被识别为异常的凝血系统。对处于静脉血栓形成风险中的个体的识别是能够受益于创新性技术方法的研究领域。

[0004] 与血栓形成的患者特异性风险有关的不确定性导致不接受抗凝处置的处于(血栓形成的)高风险中的患者的不必要的血栓形成。另一方面，该不确定性导致接受不必要的抗凝处置的处于相对低风险中的患者的出血。大多数常规临床决策支持系统适于基于许多临床风险因子来估计血栓形成风险。诸如固定和避孕药使用的许多临床风险因子(对于没有明显基因缺陷的患者)已经被识别出，但是这些不足以用于筛查目的。在实践中，如在Durieux等人的“A Clinical Decision Support System for Prevention of Venous Thromboembolism”中所描述的，使用基于临床风险因子的指南。在US 2009/0298103 A1中公开了与基于分层的临床风险因子相比在概念上不同的领域，其中，基于蛋白质的单个模拟的测量结果(即，凝血酶生成测定)被链接到血栓形成风险。US 2009/0089079公开了一种基于对象的临床风险因子来确定是否发出用于考虑关于对象的血栓形成风险的预防的警报的方法。Penco等人的“Assessment of the role of genetic polymorphism in venous thrombosis through artificial neural networks”(Annals of Human Genetics，第69卷，第693至706页(2005))公开了一项利用人工神经网络的属性来探索在静脉血栓形成事件与多位点基因型之间的关联的研究。Haan等人的“Multiple SNP testing improves risk prediction of first venous thrombosis”(Blood，第120卷，第656至663页(2012))公开了一种用于基于基因标记物与临床风险因子的组合来预测第一静脉血栓形成的风险的方法。Novis等人的“Prevention of thromboembolic events in surgical patients

through the creation and implementation of a computerized risk assessment program"(Journal of Vascular Surgery,第51卷,第648至654页(2010))公开了一种用于基于临床风险因子来防止血栓形成的发展的方法。然而,以上方法并不足够特异以用于血栓形成的筛查,因为使用当前可用的方法而被错误分类的患者的数量仍然很高。

发明内容

[0005] 本发明的目的是提供一种对于具有个人特异性的血栓形成风险估计具有增加的准确性的临床决策支持系统。

[0006] 该目的通过如权利要求1中所要求保护的装置、如权利要求9中所要求保护的方法以及通过如权利要求15所要求保护的计算机程序产品来实现。

[0007] 因此,对临床风险因子和分子标记物的两个概念上不同的领域进行组合。所提出的组合是有必要进行的并且需要机器学习和数据驱动方法的显著努力。风险因子和蛋白质浓度的最小集一起使针对血栓形成风险的最佳预测值被选择并且使数值算法被创建,所述数值算法将选定的因子和浓度的数值转化为指定血栓形成风险的单个数值。由此,能够基本上增加个人特异性血栓形成风险估计的准确性,尤其是在具有至少一个已知临床风险因子存在的患者的增加的风险子组内。该子组(尤其)涉及住院的、妊娠的或者(开始)使用口服避孕药并且因此受到医师的注意的患者。在本上下文中,所提出的解决方案辅助医师针对已知增加血栓形成风险的状况来将处置的或检查的患者分层为高风险类别和低风险类别。具体地,所提出的解决方案可以被用于基于估计的血栓形成风险来决定每个患者是否要施予抗凝血处置。

[0008] 此处术语"分子标记物"旨在包括作为患者表现型的指示器的诸如蛋白质或多核苷酸的生物分子或生物分子的部分的存在或浓度。这样的存在或浓度可以直接在例如血液或组织样本中被测得或作为诸如实时定量聚合酶链式反应(PCR)或凝血酶基因测定的功能性测试中的分子的测量结果。

[0009] 根据第一方面,至少一个蛋白质浓度可以是从以下中选择的:血液中的凝血蛋白FVIII的浓度、血液中的凝血蛋白FXI的浓度、以及血液中的凝血蛋白TFPI的浓度。基于从临床研究获得的患者数据集,这些类型的蛋白质浓度已经被证明是用作血栓形成风险的可靠指示器。

[0010] 根据能够与上面的第一方面组合的第二方面,至少一个临床风险因子可以是从以下中选择的:在第一预定时间段内的固定、在第二预定时间段内的外科手术、静脉血栓形成的家族史、在第三预定时间段内的妊娠或分娩、对雌激素的当前使用以及肥胖。在具体范例中,所述第一预定时间段可以对应于至少三个月,所述第二预定时间段可以对应于一个月,并且所述第三预定时间段可以对应于至少三个月。这些临床风险因子已经基于上面的具体临床研究的患者数据集被选择作为与上面的具体蛋白质浓度相组合最为可靠的。

[0011] 根据能够与上面的第一方面或第二方面组合的第三方面,可以将血栓形成风险的所述估计值与预定阈值进行比较以便基于所述比较结果对所述估计值进行分类。由此,可以通过将患者分类为预定的风险水平组,例如高血栓形成风险组和低血栓形成风险组,来支持由临床医师做出的决策。

[0012] 根据第三方面的具体实施方式,可以允许用户输入所述预定阈值或禁用所述预定

阈值。由此，能够基于用户(即，临床医师)的需求对所述决策支持机构进行调整。

[0013] 根据能够与上面的第一方面至第三方面中的任一方面组合的第四方面，可以提供用于基于在数据库中存储的数据集通过优化流程来应用学习过程的优化机构，从而使预测误差最小化。这允许所述临床决策支持机构到新患者的新数据集或到个体患者的具体数据集的持续调整。

[0014] 根据第四方面的具体实施方式，所述数据集可以被划分为训练集、验证集以及测试集，其中，所述训练集和所述验证集可以被用于选择机器学习函数的类型和用于优化分类器的模型参数集，其中，经优化的分类器可以被用于获得所述患者特异性输入特征，并且其中，所述测试集可以被用于基于所获得的输入特征来监测针对所述测试集的患者的所述估计值。

[0015] 根据另一实施例，所述处理器适于基于临床风险因子、单核苷酸多态性(SNP)以及蛋白质水平来计算表示患者的血栓形成风险的估计值的深静脉血栓形成(DVT)风险得分。

[0016] 应注意，所述装置可以被实施为具有离散硬件部件的离散硬件电路、被实施为集成芯片、被实施为芯片模块的布置或者被实施为由软件例程或程序控制的信号处理设备或芯片，所述软件例程或程序被存储在存储器中、被编写在计算机可读介质上或是从诸如因特网的网络下载的。

[0017] 应理解，根据权利要求1所述的装置、根据权利要求9所述的方法以及根据权利要求15所述的计算机程序产品具有相似和/或相同的优选实施例，尤其是如在从属权利要求中定义的实施例。

[0018] 应理解，本发明的优选实施例也能够是从属权利要求与各自的独立权利要求的任何组合。

[0019] 本发明的这些方面和其他方面将从下文描述的实施例变得显而易见并将参考下文描述的实施例得到阐述。

附图说明

[0020] 在附图中：

[0021] 图1示出了根据各个实施例的临床决策支持系统的示意性框图；

[0022] 图2示出了根据第一实施例的风险估计流程的流程图；

[0023] 图3示出了根据第二实施例的分类器优化流程的流程图；

[0024] 图4示出了根据第三实施例的用户接口的示意性表示；

[0025] 图5A和5B分别示出了由支持向量机仅利用临床风险因子作为输入预测的血栓形成的受试者工作特性曲线(ROC)加95％置信区间，以及由分类器利用临床风险因子和蛋白质浓度作为输入预测的血栓形成的ROC曲线加95％置信区间；以及

[0026] 图6A和6B分别示出了由支持向量机仅利用临床风险因子作为输入在具有一个或多个已知临床风险因子存在的患者的子组内预测的血栓形成的ROC加95％置信区间，以及由分类器利用临床风险因子和蛋白质浓度作为输入预测的血栓形成的ROC曲线加95％置信区间。

具体实施方式

[0027] 现在基于计算机化的临床决策支持系统来描述实施例，所述计算机化的临床决策支持系统用于基于对临床风险因子与诸如蛋白质浓度的分子标记物的组合的考虑来预测血栓形成风险。

[0028] 图1示出了根据各个实施例的临床决策支持系统的示意性框图，所述临床决策支持系统涉及临床决策支持算法和/或软件。临床决策支持系统包括：数据接口(DI)10，其中，使与具体患者有关的信息对所述系统可用；处理器(P)20，其应用解释算法；以及用户接口(UI)30，其使对计算的数据的解释对诸如临床医师的用户可用。此外，任选的优化系统可以被提供用于优化分类器，从而提供在良好预测准确性与用于临床决策支持算法的输入特征或参数集的简洁性之间的良好权衡。优化系统包括优化单元(O)40，所述优化单元可以是基于运行优化软件的单独处理器的或基于控制处理器20的单独软件例程的。优化单元40从数据库(DB)50检索优化所需的数据。

[0029] 数据接口10可以是用于允许用户与临床决策支持系统之间的交互的经典用户接口，或者到中央计算机数据库或电子病历的直接链接。在任一种情况下，数据接口10适于在临床决策支持系统用于评估血栓形成风险的日期时收集关于患者的以下输入特征中的至少一些：

[0030] 最近三个月内的固定(石膏绷带，在家卧床休息长达至少4天，住院，例如，“1”代表真，“0”代表假)；

[0031] 在上个月内的外科手术(例如，“1”代表真，“0”代表假)；

[0032] 静脉血栓形成的家族史(在至少一个父母亲、兄弟或者姐妹经历了静脉血栓形成时被认为是阳性的(例如，“1”代表真，“0”代表假))；

[0033] 最近三个月内的妊娠或分娩(例如，“1”代表真，“0”代表假)；

[0034] 正在使用雌激素(口服避孕药或荷尔蒙替代治疗(例如，“1”代表真，“0”代表假))；

[0035] 肥胖(身体质量指数超过30(例如，“1”代表真，“0”代表假))；

[0036] 血液中的凝血蛋白FVIII的浓度(U/ml)；

[0037] 血液中的凝血蛋白FXI的浓度(U/ml)；以及

[0038] 血液中的凝血蛋白TFPI的浓度(ng/ml)。

[0039] 在上文中，出于清晰的目的给出针对每个输入特征的单位以及可能数值，但是对具体单位的选择不是必须的。

[0040] 基于上面输入特征中的至少一些，处理器20通过应用临床决策支持算法来计算上面的数值输入列表的数值函数。该数值函数返回在零与一之间的数，即风险得分(R)，其中，零是最低可能血栓形成风险指示并且一是最高可能血栓形成风险指示。该数值输出可以直接被示出在用户接口30上和/或可以与在零与一之间的阈值(T)进行比较。如果风险得分超过阈值T，则针对其值已经被输入到计算中的患者指示抗凝血治疗。否则，预防性抗凝血治疗被指示为不建议的。对T的选择确定临床决策支持系统的敏感性与特异性之间的平衡，T在系统中能够被设置为固定值或由用户在用户接口30处调谐。T的低值将暗示着朝向对高风险的指示的偏向，这导致很少假阴性(高敏感性)但是增加假阳性的数量(低特异性或处置过度)。T的高值给出相反的效应并且趋于处置不足。对T的具体选择是诸如临床医师的用户的职责并且可以是临床研究的主题，但这里不进一步讨论。

[0041] 临床决策支持系统可以被实施为能够由需要做出与患者的抗凝血处置有关的决

策的临床医师访问的计算机(系统)上的软件应用。任选地,临床决策支持系统的软件应用可以(例如,作为插件)被集成在现有医院信息管理系统中。

[0042] 解释临床决策支持算法可以是复杂数学函数,所述复杂数学函数取上面九个输入特征的数值(或布尔值)作为输入、将这些使用在一系列非线性计算中并且返回在零与一之间的数值,其中,较高的值表示较高的血栓形成风险。数值函数包括机器学习领域中常见的分类器函数中的一个或常见的分类器函数的组合,所述分类器函数例如是神经网络函数或支持向量机或贝叶斯网络。由优化单元40基于对象,即血栓形成患者的数据库50以及针对上述九个输入参数的数值是可用的所述对象的健康控制来优化这些分类器。优化单元40的优化涉及以这样的方式调谐分类器函数的参数,即使得对数据库中的对象计算的风险得分与记录的血栓形成出现之间的关联最大化。优化过程构成需要机器学习以及数值优化领域中的很强经验和理解的显著努力。所述过程还强烈地取决于下层数据库50的质量。

[0043] 图2示出了根据第一实施例的血栓形成风险估计过程的流程图。在步骤S200中开始流程之后,在步骤S201中数据接口10在医院电子病历(EPR)存在时访问医院电子病历(EPR),并且读出上面列出的九个患者特征。任选地,可以请求或允许用户例如经由用户接口30人工输入针对不可从EPR获得的患者参数的数值。然后,在步骤202中,数据接口10检查输入的值的正确数值格式,并且在输入格式与需要的格式不匹配时能够生成错误消息。在不正确的格式的情况下,在必要时,在步骤S203中将数据转换为在上面的列表中指示的数值格式。额外地,用户接口30可以允许用户输入在零与一之间的阈值T的数值或者禁用阈值。

[0044] 然后,在步骤S204中,流程(例如,通过在用户接口30处的各自按钮上点击)检查用户是否已经请求了风险计算。如果没有,则系统重复上面的步骤S201至S203以允许对输入特征的更新,或者简单地重复步骤204直到请求了风险计算为止。即,步骤S204的“否”分支箭头能够简单地指回步骤S204的顶部并且不需要回到步骤S201。如果在步骤S204中检测到所述请求,则在步骤S205中(例如由处理器20)调用临床决策支持算法基于在先前步骤中收集的输入特征来计算风险得分。

[0045] 在随后的步骤S206中,检查阈值(T)是否已经被启用。如果没有,则流程分支到步骤S209并且在流程结束在步骤S210中之前,例如在计算机屏幕或用户接口30的其他输出介质上将计算的风险得分示出为数字或另一图形表示。否则,如果在步骤S206中流程检测到阈值已经被启用,则在步骤S207中将风险得分与阈值进行比较并且基于比较的结果对风险得分进行分类。最后在步骤208中,取决于风险得分高于或低于阈值,使对“高血栓形成风险”或“低血栓形成风险”的分类例如在用户接口30的屏幕上可见。任选地,阈值与风险得分之间的数值和/或图形比较应当与分类一起被示出。

[0046] 根据对第一实施例的修改,能够连续地(代替在请求时)计算风险得分。这还能够利用一些缺失的输入参数来完成。在那种情况下,基于在计算中的不确定性,将(例如,由最小风险估计和最大风险估计所指示的)可能风险得分的范围提供为输出。

[0047] 在下文中,基于第二实施例来描述对临床决策支持算法的优化。

[0048] 可以基于关于针对静脉血栓形成的许多潜在风险的广泛问卷从数据集合导出数据库50的请求的数据集。更具体而言,数据集合可以涉及从问卷和如在各自的测定协议中描述的临床测定(例如,蛋白质浓度的活性或基于凝血的测定)获得的信息(例如,临床风险

因子)。

[0049] 机器学习方法是利用能够可能在数据的数值中隐藏的模式来预测输出的黑盒方法。每个方法构建数学函数,所述数学函数取观测到的数量(例如蛋白质浓度)和质量(例如固定)作为输入,并且产生预测特定期望特征的输出。通过其结构(例如,神经网络函数)和函数参数的数值(例如,神经网络中的权值)来定义这样的函数。函数结构、参数值以及数值输入的组合产生输出特征,所述输出特征可以是二值的(例如,血栓形成对无血栓形成)或者连续的(例如,血栓形成的概率)。在第二实施例中使用的具体类型的方法是支持向量机(SVM),支持向量机是机器学习领域中常用的方法(参见例如"An Introduction to Support Vector Machines and Other Kernel-based Learning Methods"(Cambridge University Press,2000)以获得更多细节)。一般在不涉及各个输入的特性(例如,生物学意义)的情况下,从数据直接"学习"隐藏的模式。学习前进通过优化流程,在优化流程中,预测误差(即,预测的模型输出与观测之间的差异的特定数值度量)被最小化。存在全部涉及数学函数的参数的变型的许多优化或误差最小化例程,以找到产生最低预测误差的参数值集。存在关于机器学习技术和优化方法的广泛的文献。关于更深入的见解,参考Kuncheva的"Combining Pattern Classifiers:Methods and Algorithms"(Wiley-Blackwell,2004)。

[0050] 图3示出了根据第二实施例的优化过程的流程图。

[0051] 分类器是黑盒模型的具体类别,其输出是数据元素的类别或标签,其中,每个元素由许多数值特征来描述。在本实施例中的数据元素是通过测量或既往病例已知针对其的许多临床特征的人类对象。所述类别是二元的:血栓形成患者或控制对象。在包含每个参与者的数值特征和对应标签的数据库50的数据集上训练分类器。

[0052] 在步骤S300中开始优化流程之后,在步骤S301中,数据库50的数据集被划分为三个大小相等的集,称为训练集、验证集和测试集,其每个包含相同的病例对控制比率。在步骤302中,训练集用于训练或参数调谐,即,寻找使所述预测或在这种情况下使分类误差最小化的参数值集。大多数机器学习方法遭受所谓的"过度拟合",其中,所述方法在训练集上的性能比其在未用于训练的新数据上的性能好得多。因此,在步骤S303中,单独的验证集被用于测试是否出现了这样的过度拟合。训练数据和验证数据的组合允许找到机器学习函数的类型和能够攫取隐藏在(训练)数据中的真实模式又总体上仍然足以在单独的验证数据并因此也在未来的数据上预测良好的模型参数的选择。在步骤S304中,经由此优化的分类器被用于在测试集中的患者中的每个上进行预测,所述测试集在前面的优化步骤中仍然未被使用。该预测的质量(例如,在敏感性和特异性方面)是对选择的分类器的有效性的最终测试。随机地选择测试集以获得可靠的统计值。

[0053] 步骤S301至S303描述了基于数据库的训练子集和验证子集对优化分类器的选择。在步骤S305中,通过对训练集和验证集中的对象的置换(交换所述两个集中的患者),能够创建分类器的系综,每个分类器对应于训练对象和验证对象的一个具体置换。这样的系综用作投票系统。这意味着系综中的每个分类器将标签分配给相同的对象,例如,"控制对象"或"血栓形成患者"。最常出现的标签被假设为是正确标签,并且支持该标签的投票的分数用作置信度得分:如果系综中的全部分类器投票给血栓形成,则100%地确定所述参与者将罹患血栓形成,然而投票的五十-五十的分布使所述分类几乎等于抛硬币。将风险得分(R)与阈值(T)进行比较,其中,超过阈值的得分指示病例,而低于阈值的得分指示控制对象。

[0054] 当已经在步骤S305中找到在完全特征集上的最优分类器时，在步骤S306中，分析分类器中的每个输入特征的相对重要性。在训练集和验证集中选择的对象现在被用于选择对正确分类贡献最多的那些特征。为了实现这个目的，在步骤S306中针对经优化的分类器中的每个运行以下输入精简流程：

[0055] 对于到分类器的每个输入特征i，

[0056] 移除输入特征i，

[0057] 在训练集上重新优化精简的分类器，

[0058] 计算在训练集上得到的预测误差，

[0059] 还原输入特征i，

[0060] 永久地移除具有最低预测误差的输入特征，

[0061] 从开始重复直到仅保留一个输入特征为止。

[0062] 当分类器中的输入特征的数量减少时，预测误差增大。因此，总是存在在良好预测能力与使用的输入特征集的简洁性之间的权衡。上面的精简流程被用于推断对全部最具预测性的特征的选择。针对将完整数据库到训练集、验证集和测试集的每个上述(随机)划分执行所述精简流程。在步骤S307中，对于每个划分，分类器被精简到十个输入特征，并且每个剩余的输入特征被标记。然后在步骤S309中，对每个输入特征保留在“前十个”中的次数进行计数，并且该计数被用于将输入特征从最具预测性(最常见的前十个的部分)到最不具预测性进行排序。最后，最具预测性的输入特征被用于在处理器20的临床决策支持算法中的风险计算，并且所述流程结束在步骤S310中。

[0063] 因此，第二实施例的优化流程能够被用于基于数据库50中的新患者数据定期地更新处理器20的临床决策支持系统。

[0064] 图4示出了图1的用户接口30的前视图的示意性表示。在左侧部分中，患者姓名(PN)和其身份标号(ID)被指示为“Jane Doe”和“099812”。在该信息下面，选定了九个输入特征，并且上面的患者的所述特征的实际二元值(“0”或“1”)被指示在名称下面的右侧上。前六个输入特征是指示以下的临床风险因子：近期的外科手术(RS)、肥胖(O)、家族史(FH)、固定(I)、避孕药使用(CU)以及妊娠(P)。后三个输入特征是以下的浓度水平：凝血蛋白因子VIII(FVIII)、因子XI(FXI)以及组织因子通道抑制剂(TFPI)。在右侧部分上，指示了当前设置的阈值水平(T)(即，0.5)并且在下面指示了禁用(DA)功能的状态。这可以简单地是光或颜色指示器。再往下面，示出了用于通过处理器20激活或触发风险计算的按钮(CAL)。在该按钮下面，提供了计算的风险得分(RS)的数值指示(即，0.12)，并且再往下面，该风险在风险尺度上相对于阈值T的图形表示(RV)被示为分层。在风险尺度上指示当前风险得分的条被量化为低风险(LR)。该可视化与其他输出信息以及在用户接口30上的输入功能一起允许用户，即临床医师的快速评估并且提供针对处置决策的增强的支持。

[0065] 通过对本发明的说明的方式来呈现下面的范例，而不旨在以任何方式限制本发明以及本文提供的实施例。

[0066] 在涉及血栓形成风险分类的第一范例中，上面说明的第二实施例被应用于～500血栓形成患者和～500健康控制的临床研究，并且示出所提出的解决方案在估计准确性方面比单单基于临床风险因子的“常规”方法得到更好的结果。支持向量机的系综被用在LeidenThrombophilia研究(LETS)(如例如在van der Meer等人的“The

LeidenThrombophilia Study(LETS)”(Thromb Haemost,第78卷,第1号,第631至635页(1997))中描述的)上,以便找到能够区分血栓形成患者与健康控制的已知生物标记物的组合。焦点指向两种不同类型的患者特征,即血液中的凝血蛋白浓度和已知与血栓形成有关的临床风险因子。能够示出,单单临床风险因子作为简单风险因子计数或者被用在机器学习方法中的预测能力能够通过并入测得的凝血蛋白浓度来改进。

[0067] 图5A和5B示出了关于以下的各自的图:由支持向量机仅利用临床风险因子作为输入预测的结果为0.72(0.68-0.77)的ROC曲线下面积(AUC)的血栓形成的受试者工作特性曲线(ROC)加95%置信区间(图5A),以及由分类器利用临床风险因子和蛋白质浓度作为输入预测的结果为0.78(0.74-0.83)的ROC曲线下面积(AUC)的血栓形成的ROC曲线加95%置信区间(图5B)。ROC曲线绘制针对不同阈值的真阳性率(纵轴)对假阳性率(横轴)。ROC曲线下面积(AUC)被用作针对分类器系综的质量的度量。如能够从图5A和5B推断的,两种类型的特征的组合给出明显更好的分类(即,0.78对0.72的AUC,$p<0.001$)。

[0068] 第二范例涉及输入特征精简。在研究中,在凝血分类中确定的最具影响力的蛋白质是凝血因子VIII,接着是因子XI和TFPI(参见下面的表1)。利用全部临床风险因子(并不必须要针对其的测量)以及这三个蛋白质浓度的分类实现在0.77的AUC的几乎等价的分类。改进在增加的风险群体中尤其清楚,所述增加的风险群体此处定义为示出一个或多个已知临床风险因子的那些对象。

[0069] 图6A和6B示出了由支持向量机仅利用临床风险因子作为输入在具有一个或多个已知临床风险因子存在的患者的子组内预测的结果为0.67(0.60-0.75)的AUC的血栓形成的ROC加95%置信区间(图6A),以及由分类器利用临床风险因子和蛋白质浓度作为输入预测的结果为0.75(0.69-0.81)的AUC的血栓形成的ROC曲线加95%置信区间(图6B)。

[0070] 如能够从图6A和6B推断的,对三个蛋白质浓度值的使用允许具有0.75的ROC得分的该风险组比具有基于单单使用临床风险因子的0.67的ROC得分的该风险组进一步分层(共同出现因子的数量或存在哪个因子的知识)。

[0071] 表1示出了分类器特征的列表,通过保留最后删除的10个特征中的特征的分类器(基于对验证集的不同随机选择)的百分比对所述分类器特征进行排序。

[0072]

顺序	特征名称	分类器(%)	顺序	特征名称	分类器(%)
1	F8	100	14	肥胖	23
2	避孕药使用	100	15	蛋白质C	21
3	固定	100	16	F9	17
4	外科手术	100	17	蛋白质S	14
5	血栓形成的家族史	89	18	ZPI	12
6	F11	80	19	F13	8

[0073]

7	妊娠/分娩	74	20	F2	7
8	TFPI	74	21	AT	5
9	C4BP	50	22	PCI	2
10	蛋白质Z	37	23	F10	1

11	F12	37	24	F7	0
12	纤维蛋白原	26	25	F5	0
13	TAFI	24			

[0074] 表1

[0075] 已经通过使用来自MEGA(对静脉血栓形成的风险因子的多环境和基因评估)研究和Leiden Thrombophilia研究(LETS)评价了深静脉血栓形成的风险。两者都是已经在荷兰(Blom(2005);van der Meer FJ、Koster T、Vandenbroucke JP、Briët E(1997))执行的被设置为识别静脉血栓形成的风险因子的病例-控制研究。已经从具有静脉血栓形成的患者和控制获取了从凝血蛋白水平到环境血栓形成风险因子的范围的许多变量。出于该研究的目的,神经网络方法(参见例如Kuncheva(2004))已经被用在MEGA研究中以估计深静脉血栓形成(DVT)的潜在风险因子以及其在一种集成方法中的预测值。在MEGA研究上采用内部交叉验证并且在LETS研究上采用独立验证对识别的组合风险得分进行验证。

[0076] 在过去已经示出,临床风险因子与单核苷酸多态性(SNP)的组合允许在高风险患者与低风险患者之间进行辨别,其中,在MEGA上具有0.82的受试者工作特性(ROC)曲线下面积(AUC)并且在LETS上具有0.77的受试者工作特性(ROC)曲线下面积(AUC)。现在示出,通过增加蛋白质水平作为预测因子,能够实现预测准确性的明显进一步增加,如分别采用0.87和0.81的AUC量化的。

[0077] 另外,现在考虑了不可用于初始研究的四个临床风险因子:由于石膏绷带的固定、在过去3个月中的腿部损伤、从指示日期的五年前到六个月后的时间段中的癌症以及过去2个月中的超过四小时的旅行。其他考虑的风险因子也是初始研究的部分:由于长达至少4天的在家卧床休息的固定、住院、外科手术、静脉血栓形成的家族史(在至少一个父母亲、兄弟、姐妹经历了静脉血栓形成时被认为是阳性的)、在指示日期之前3个月内的妊娠或分娩、在指示日期时对雌激素的使用(口服避孕药或荷尔蒙替代治疗)以及肥胖的存在(被确定为身体质量指数为30kg/m2或更高)。

[0078] 除了来自问卷的数据和测得的蛋白质水平,数据在存在五个基因方面的情况下是可用的,所述五个基因方面即血型、以及在F2(G20210A)、纤维蛋白原(rs no 2066865)、F11(rs no 2036914)以及F5(FV Leiden;rs no 6025)中的四个单核苷酸多态性。所述数据还包括每SNP受影响的等位基因的数量。

[0079] 考虑的蛋白质水平是之前包含的蛋白质的子集(由于在MEGA研究中执行的测量结果的更有限的设置)。它们是:抗凝血酶(AT)、凝血酶原(因子II)、因子7(FVII)、FVIII、FIX、FX、FXI、纤维蛋白原以及蛋白质C(全部活性测量结果)和蛋白质S(抗原测量结果)。

[0080] 关于MEGA的交叉验证结果。考虑了基于临床风险因子、基因效应以及蛋白质水平预测风险的基于神经网络的风险得分到基于临床风险因子和基因效应(没有蛋白质水平)的风险得分以及仅基于临床风险因子的临床风险得分。所述比较被执行在MEGA研究上,但是在其他方面采用相同的交叉验证设置并且利用如在初始研究中描述的相同方法。对应的AUC是0.87、0.83以及0.78,即每次增加改进风险得分的准确性;全部改进是显著的(在配对t检验中$p<0.01$)。

[0081] LETS研究比MEGA研究少包括四个临床风险因子,如上面关于临床风险因子所描述的。已经在没有这四个风险因子的情况下并且在排除癌症患者的情况下重复了如在前段中

执行的交叉验证，所述癌症患者也已经从LETS研究排除。在精简的MEGA研究上的AUC是0.84、0.80和0.74，采用与上段中相同的顺序。接下来，对于输入特征(临床风险因子、具有/没有基因效应、具有/没有蛋白质水平)的选择中的每个，导出在精简的MEGA研究上的一个风险得分(没有如将在交叉验证中必须的到训练集和测试集的划分)，并且在没有对LETS研究的个体的调整的情况下应用该风险得分。得到的AUC是0.82、0.79和0.74，从而示出所提出的风险得分能够在几乎没有损失性能的情况下被应用在独立研究上，并且由于所提出的对蛋白质水平的包括的改进保持在外部验证中。

[0082] 相同的方法被使用在具有上述临床风险因子中的一个或多个存在的个体的MEGA子群体内的交叉验证研究中(这对于LETS研究而言也在初始归档中完成)。针对三种得分方法得到的AUC是0.86、0.81和0.76，同样，其中，较低的得分代表考虑较少输入特征的得分。

[0083] 在如上面描述的相同的方法之后，对用作对提供风险得分的神经网络的输入的全部特征的重要性进行排序。结果被示出在表2中。所述结果与早前的结果部分重叠：F8目前仍然是最具预测性的蛋白质，并且避孕药使用、外科手术、固定以及家族史仍然得分很高。TFPI在MEGA中未被测量并且因此不出现在排序中。F11比之前的得分低得多。

[0084]

顺序	特征名称	前10(%)	顺序	特征名称	前10(%)
1	F8	100	14	F11SNP	15
2	口服避孕药使用	100	15	凝血蛋白	13
3	腿部损伤	100	16	蛋白质C	13
4	FV Leiden	100	17	妊娠	12
5	外科手术	88	18	血型	12
6	固定(住院)	87	19	AT	10
7	家族史	85	20	FIX	9
8	蛋白质S	68	21	FXI	8
9	纤维蛋白原SNP	54	22	石膏绷带	8
10	固定(在家)	38	23	癌症	5

[0085]

11	肥胖	22	24	FVII	5
12	FX	21	25	纤维蛋白原	5
13	F2SNP	17	26	旅行	3

[0086] 表2

[0087] 在具有基于全部临床风险因子、一个SNP(FV Leiden)以及FVIII的蛋白质水平的风险得分的在MEGA上的交叉验证提供了仅降低了少许的准确性(AUC＝0.85对0.87)。对纤维蛋白原中的SNP以及蛋白质S和FX的蛋白质水平的进一步增加将AUC增加到0.86。

[0088] 如上面说明的，基于临床风险因子、SNP以及蛋白质水平的DVT风险得分比在MEGA研究上的评价中没有蛋白质水平的已知方法在敏感性/特异性方面示出了明显改进。

[0089] 总而言之，已经描述了一种用于临床决策支持的装置和方法，所述装置和方法用于基于临床风险因子和诸如蛋白质浓度的分子标记物的组合来识别处于高血栓形成风险中的患者。这些临床风险因子和分子标记物被组合在基于机器学习的算法中，所述基于机

器学习的算法返回与未来的血栓形成事件的估计的风险相关的输出值。

[0090] 尽管已经在附图和前面的描述中详细说明和描述了本发明，但是这样的说明和描述被认为是说明性或示范性的而非限制性的。本发明不限于所公开的实施例。本发明能够应用于任何临床决策支持领域中，应用于其中需要做出与是否要将患者放置于预防性治疗下有关的决策的情形。此外，输入特征(即，临床风险因子和分子标记物)的数量和类型不限于实施例中提到的九个输入因子。基于上面的范例的优化流程，各种其他临床风险因子或分子标记物(例如，蛋白质Z、C4B结合蛋白、纤维蛋白原、TAFI、因子II、V、VII、IX、X、XII或XIII、抗凝血酶、蛋白质C、蛋白质C抑制剂、蛋白质S或其他标记物的浓度)可以被选择作为决策输入特征。

[0091] 通过研究附图、公开内容以及权利要求书，本领域技术人员在实践要求保护的本发明时能够理解并实现所公开实施例的其他变型。在权利要求书中，“包括”一词不排除其他元件或步骤，词语“一”或“一个”不排除多个。单个处理器或其他单元可以实现权利要求中记载的若干项目的功能。在互不相同的从属权利要求中记载特定措施并不指示不能有利地使用这些措施的组合。

[0092] 前面的描述详细描述了本发明的特定实施例。然而应认识到，无论前面如何详细地出现在文本中，本发明可以以多种方式来实现，并且因此不限于公开的实施例。应当指出，当描述本发明的特定特征或方面时对特定术语的使用不应被认为意指术语在本文中被重新限定为限于包括本发明的与该术语相关联的特征或方面的任何具体特性。

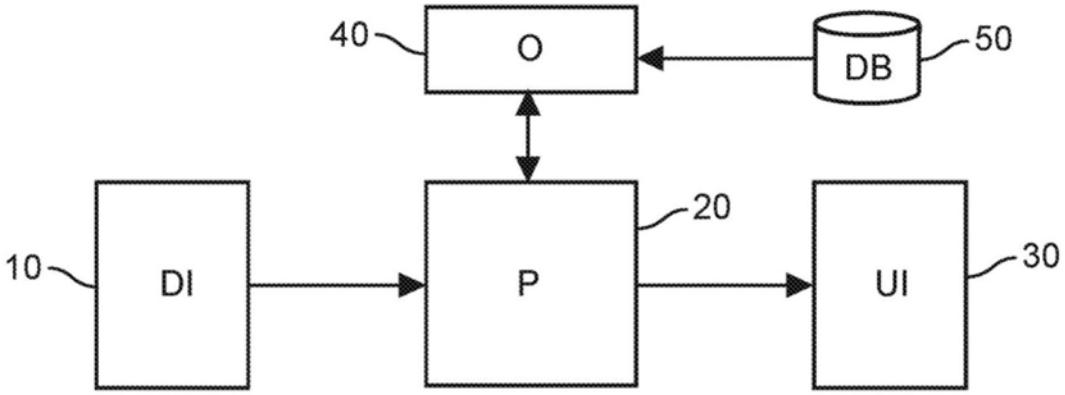

40
O
DB
50
10
DI
P
20
UI
30

图1

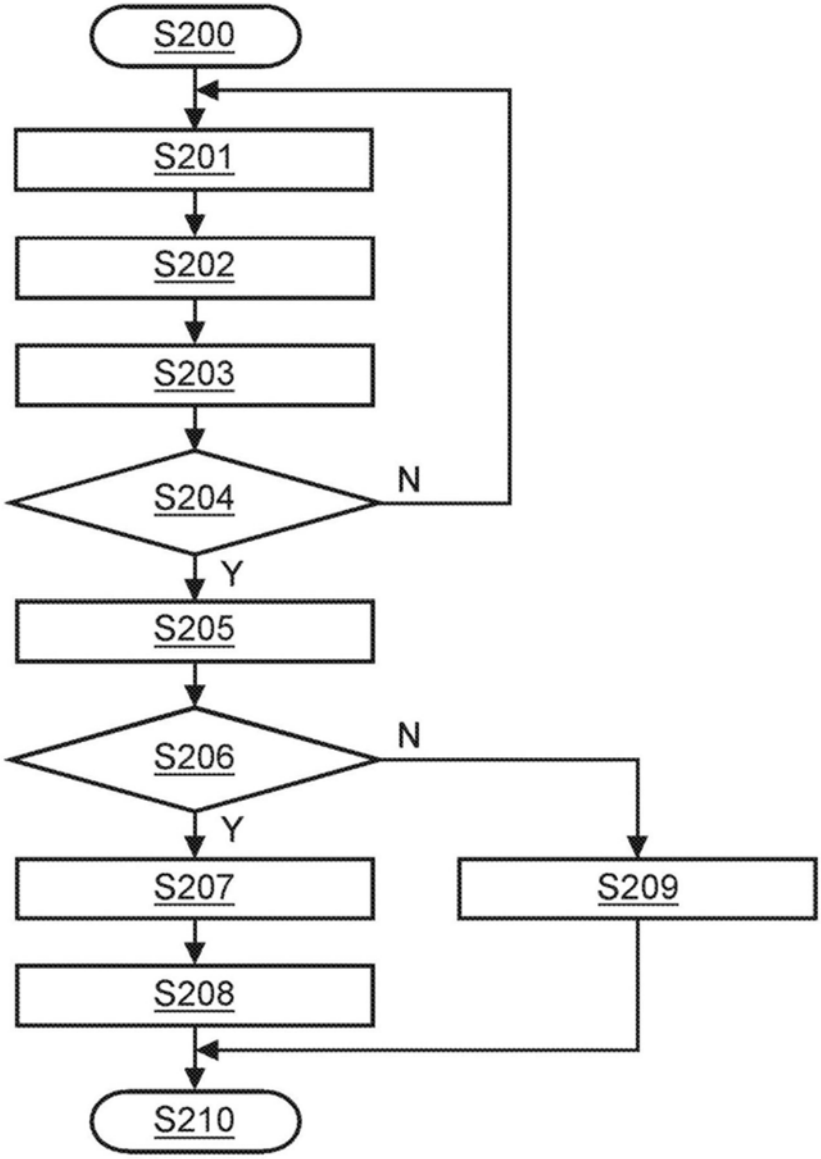

S200
S201
S202
S203
S204
N
Y
S205
S206
N
Y
S207
S209
S208
S210

图2

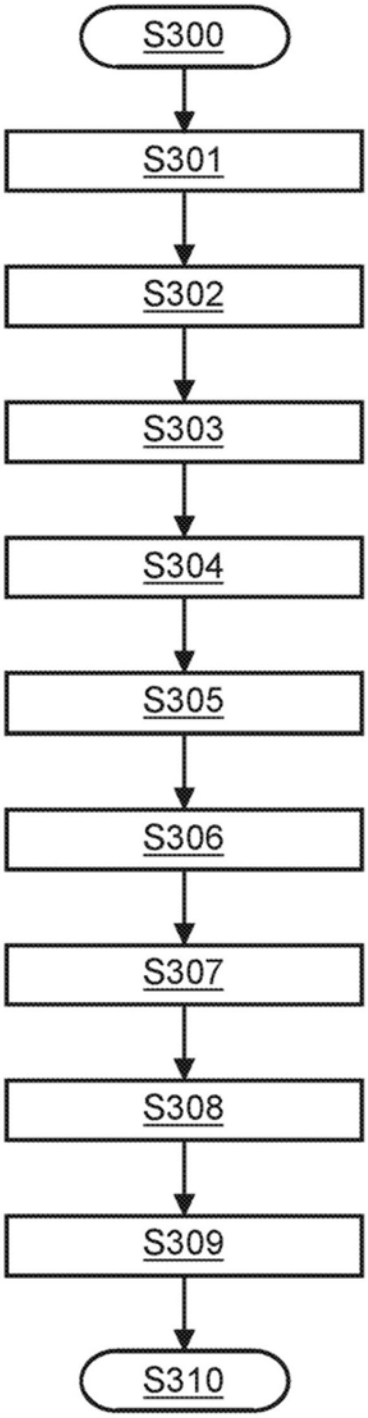


图3

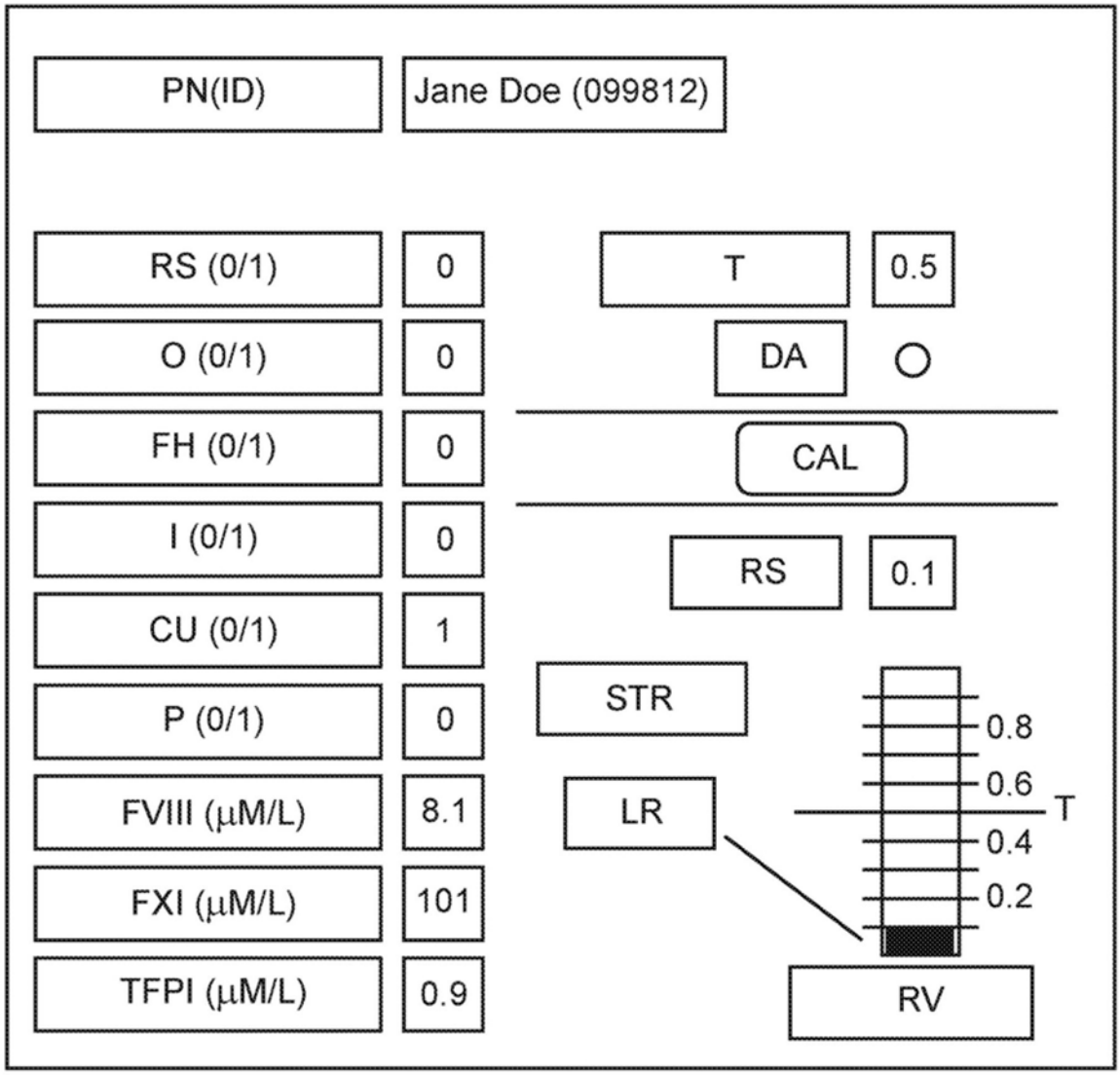

PN(ID)
Jane Doe (099812)
RS (0/1)
0
O (0/1)
0
FH (0/1)
0
I (0/1)
0
CU (0/1)
1
P (0/1)
0
FVIII (μM/L)
8.1
FXI (μM/L)
101
TFPI (μM/L)
0.9
T
0.5
DA
CAL
RS
0.1
STR
LR
0.8
0.6
T
0.4
0.2
RV

图4

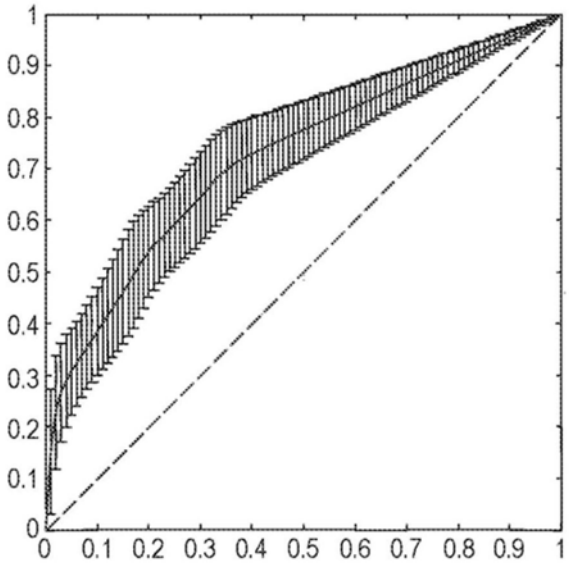


图5A

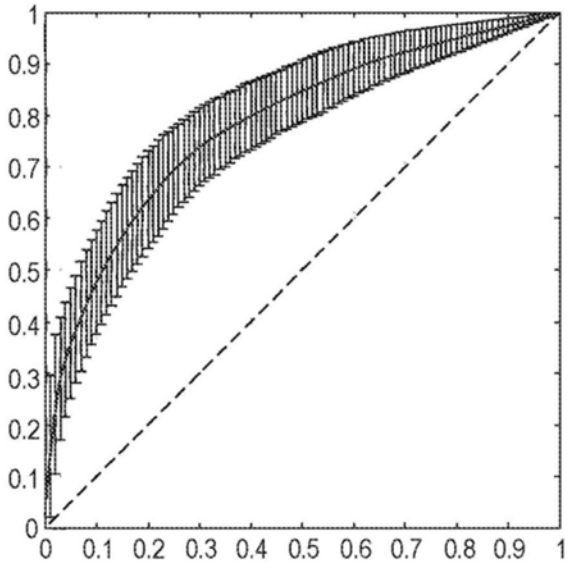


图5B

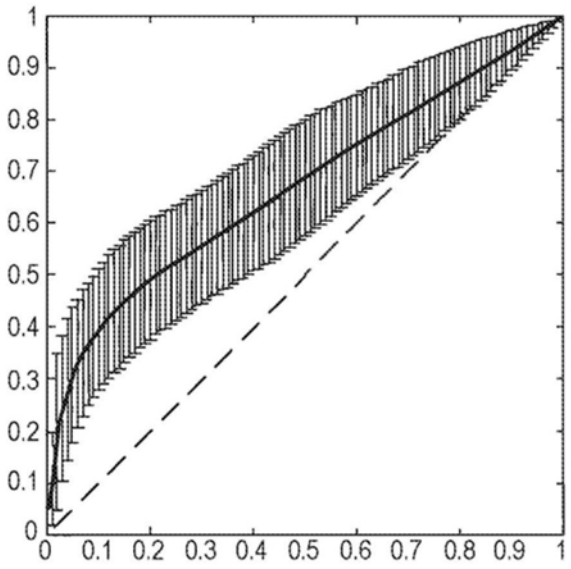

1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

图6A

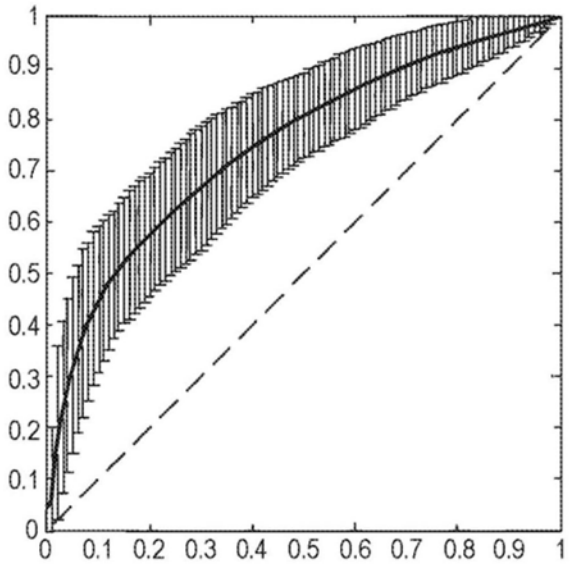

1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

图6B