(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 114868204 A
(43)申请公布日 2022.08.05

(21)申请号 202080085251.8

(22)申请日 2020.12.09

(30)优先权数据
20315299.6 2020.06.09 EP
62/945,814 2019.12.09 US

(85)PCT国际申请进入国家阶段日
2022.06.08

(86)PCT国际申请的申请数据
PCT/US2020/064100 2020.12.09

(87)PCT国际申请的公布数据
WO2021/119188 EN 2021.06.17

(71)申请人 赛诺菲
地址 法国巴黎

(72)发明人 A-G·拉德纳克 P·布莱斯
E·德雷纳尔迪斯
R·A·赫尔南德斯-维西诺
F·J·J·希门尼斯
C·M·莫洛尼

(74)专利代理机构 北京坤瑞律师事务所 11494
专利代理师 封新琴

(51)Int.Cl.
G16H 50/70(2006.01)

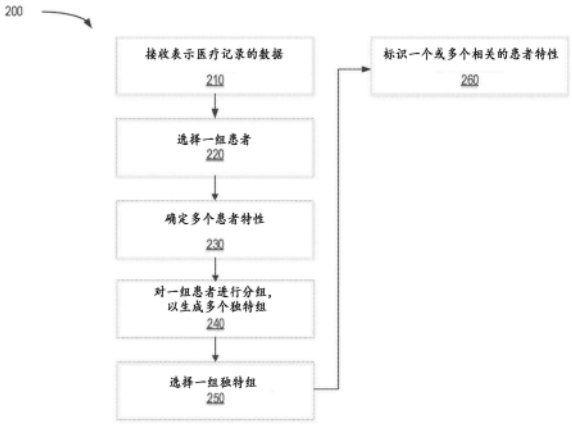
权利要求书3页 说明书18页 附图5页

(54)发明名称

用于再利用药物的数据处理系统和方法

(57)摘要

用于再利用药物的数据处理系统可以包括：计算机可读存储器，所述计算机可读存储器包括计算机可执行指令；以及至少一个处理器，所述至少一个处理器被配置成执行包括所述计算机可执行指令和至少一个机器学习模型的可执行逻辑以执行一个或多个操作。所述一个或多个操作可以包括：接收表示多个患者的医疗记录的数据；选择一组患者；确定所述一组患者的多个患者特性；根据所述多个患者特性对所述一组患者进行分组，以生成多个独特组，所述独特组中的每个独特组包括所述一组患者中的至少一个患者；基于一个或多个组选择标准选择一组独特组；以及通过分析所述一组独特组中的每个独特组来标识一个或多个相关的患者特性。



CN 114868204 A

1.一种用于再利用药物的计算机实现的方法,所述计算机实现的方法包括:

由计算机系统接收表示多个患者的医疗记录的数据;

通过以下基于所述医疗记录选择一组患者:

确定至少一个与所述药物相关联的目标信号传导途径;以及

基于对应于与所述目标信号传导途径联系的诊断的一个或多个因素确定一个或多个指标;

确定所述一组患者的多个患者特性,所述一组患者中的每个患者表现出所述多个患者特性中的至少一个患者特性;

由所述计算机系统根据所述多个患者特性对所述一组患者进行分组,以生成多个独特组,所述独特组中的每个独特组包括所述一组患者中的至少一个患者;

基于一个或多个组选择标准选择所述多个独特组中的一组独特组;以及

通过分析所述一组独特组中的每个独特组来标识一个或多个相关的患者特性。

2.根据权利要求1所述的方法,其中对所述一组患者进行分组包括:

执行被配置成执行一个或多个无监督聚类技术的机器学习系统。

3.根据权利要求2所述的方法,其中所述一个或多个无监督聚类技术包括二分k均值聚类技术。

4.根据权利要求1-3中任一项所述的方法,其中对所述一组患者进行分组包括执行多重对应分析以减少所述多个患者特性的维度。

5.根据权利要求1-4中任一项所述的方法,其中选择所述一组独特组包括:

针对所述多个独特组中的每个独特组,确定由所述独特组表现出的每个患者特性的特征得分;以及

将所述多个独特组中的每个独特组的所述特征得分与特征得分阈值进行比较。

6.根据权利要求1所述的方法,其中选择一组独特组包括以下中的至少一个:确定所述多个独特组中的每个独特组的稳定性量度;确定所述多个独特组中的每个独特组的纯度量度;以及确定所述一组患者的包括在所述多个独特组中的每个独特组中的患者数量。

7.根据权利要求1所述的方法,其中标识所述一个或多个相关的患者特性包括:

针对所述一组独特组中的每个独特组,通过选择所述多个患者特性中的由所述独特组表现出的并且对应于所述组的主题的患者特性来生成多个潜在相关的患者特性;以及

对所述多个潜在相关的患者特性中的所述潜在相关的患者特性中的每一个进行评级。

8.根据权利要求7所述的方法,其中对所述潜在相关的患者特性中的每一个进行评级包括:针对每个潜在相关的患者特性,基于所述潜在相关的患者特性和至少一个与所述药物相关联的参考适应证的同现频率分配等级值。

9.根据权利要求8所述的方法,其中所述同现是通过以下测量的:针对所述潜在相关的患者特性中的每一个,确定所述一组独特组的包括所述潜在相关的患者特性和所述至少一个参考适应证两者的比例。

10.根据权利要求7所述的方法,其中标识所述一个或多个相关的患者特性包括:针对所述潜在相关的患者特性中的每一个,确定临床可行性和商业可行性中的至少一个。

11.一种用于再利用药物的数据处理系统,所述数据处理系统包括:

计算机可读存储器,所述计算机可读存储器包括计算机可执行指令;以及

至少一个处理器，所述至少一个处理器被配置成执行包括所述计算机可执行指令和至少一个机器学习模型的可执行逻辑以执行一个或多个操作，所述一个或多个操作包括：

接收表示多个患者的医疗记录的数据；

通过以下基于所述医疗记录选择一组患者：

确定至少一个与所述药物相关联的目标信号传导途径；以及

基于对应于与所述目标信号传导途径联系的诊断的一个或多个因素确定一个或多个指标；

确定所述一组患者的多个患者特性，所述一组患者中的每个患者表现出所述多个患者特性中的至少一个患者特性；

使用所述机器学习模型并且根据所述多个患者特性对所述一组患者进行分组，以生成多个独特组，所述独特组中的每个独特组包括所述一组患者中的至少一个患者；

基于一个或多个组选择标准选择所述多个独特组中的一组独特组；以及

通过分析所述一组独特组中的每个独特组来标识一个或多个相关的患者特性。

12.根据权利要求11所述的数据处理系统，其中所述机器学习模型被训练成使用一个或多个无监督聚类技术对所述一组患者进行分组。

13.根据权利要求12所述的数据处理系统，其中所述一个或多个无监督聚类技术包括二分k均值聚类技术。

14.根据权利要求11-13中任一项所述的数据处理系统，其中对所述一组患者进行分组包括执行多重对应分析以减少所述多个患者特性的维度。

15.根据权利要求11-14中任一项所述的数据处理系统，其中选择所述一组独特组包括：

针对所述多个独特组中的每个独特组，确定由所述独特组表现出的每个患者特性的特征得分；以及

将所述多个独特组中的每个独特组的所述特征得分与特征得分阈值进行比较。

16.根据权利要求11-15中任一项所述的数据处理系统，其中选择一组独特组包括以下中的至少一个：确定所述多个独特组中的每个独特组的稳定性量度；确定所述多个独特组中的每个独特组的纯度量度；以及确定所述一组患者的包括在所述多个独特组中的每个独特组中的患者数量。

17.根据权利要求11所述的数据处理系统，其中标识所述一个或多个相关的患者特性包括：

针对所述一组独特组中的每个独特组，通过选择所述多个患者特性中的由所述独特组表现出的并且对应于所述组的主题的所述患者特性来生成多个潜在相关的患者特性；以及

对所述多个潜在相关的患者特性中的所述潜在相关的患者特性中的每一个进行评级。

18.根据权利要求17所述的数据处理系统，其中对所述潜在相关的患者特性中的每一个进行评级包括：针对每个潜在相关的患者特性，基于所述潜在相关的患者特性和至少一个与所述药物相关联的参考适应证的同现频率分配等级值。

19.根据权利要求18所述的数据处理系统，其中所述同现是通过以下测量的：针对所述潜在相关的患者特性中的每一个，确定所述一组独特组的包括所述潜在相关的患者特性和所述至少一个参考适应证两者的比例。

20.一种用于再利用药物的计算机实现的方法,所述计算机实现的方法包括:

由计算机系统接收表示多个患者的医疗记录的数据;

通过以下基于所述医疗记录选择一组患者:

确定至少一个与所述药物相关联的目标信号传导途径;以及

基于对应于与所述目标信号传导途径联系的诊断的一个或多个因素确定一个或多个指标;

确定所述一组患者的多个患者特性,所述一组患者中的每个患者表现出所述多个患者特性中的至少一个患者特性;

由所述计算机系统根据所述多个患者特性对所述一组患者进行分组,以生成多个独特组,所述独特组中的每个独特组包括所述一组患者中的至少一个患者;

基于一个或多个组选择标准选择所述多个独特组中的一组独特组;以及

通过分析所述一组独特组中的每个独特组来标识一个或多个相关的患者特性;

将所述一个或多个相关的患者特性中的至少一个相关的患者特性标识为用于再利用所述药物的目标适应证。

用于再利用药物的数据处理系统和方法

优先权要求

[0001] 本申请要求于2020年6月9日提交的欧洲专利申请20315299.6和于2019年12月9日提交的美国临时专利申请序列号62/945,814的权益。前述专利申请的全部内容特此通过引用并入。

技术领域

[0002] 本公开文本总体上涉及用于药物再利用的数据处理系统和方法。

背景技术

[0003] 临床药物再利用(有时被称为重新定位)可以指一种成本可能相对较低且可以提供高效率的药物发现策略。药物再利用通常涉及分析已经被批准用于治疗一种类型的病状(例如,疾病)的药物是否可以用于治疗其他类型的病状(例如,常见和/或罕见疾病)。若干个治疗领域呈现出药物再利用的高潜力,包括肿瘤学、免疫学、感染性疾病和罕见疾病。

发明内容

[0004] 在本公开文本的至少一个方面,提供了一种数据处理系统。所述数据处理系统包括:计算机可读存储器,所述计算机可读存储器包括计算机可执行指令;以及至少一个处理器,所述至少一个处理器被配置成执行包括所述计算机可执行指令和至少一个机器学习模型的可执行逻辑。当所述至少一个处理器正在执行所述计算机可执行指令时,所述至少一个处理器被配置成执行一个或多个操作。所述一个或多个操作包括接收表示多个患者的医疗记录的数据。所述一个或多个操作包括通过以下基于所述医疗记录选择一组患者:确定至少一个与所述药物相关联的目标信号传导途径;以及基于对应于与所述目标信号传导途径联系的诊断的一个或多个因素确定一个或多个指标。所述一个或多个操作包括确定所述一组患者的多个患者特性,所述一组患者中的每个患者表现出所述多个患者特性中的至少一个患者特性。所述一个或多个操作包括使用所述机器学习模型并且根据所述多个患者特性对所述一组患者进行分组,以生成多个独特组,所述独特组中的每个独特组包括所述一组患者中的至少一个患者。所述一个或多个操作包括基于一个或多个组选择标准选择所述多个独特组中的一组独特组。所述一个或多个操作包括通过分析所述一组独特组中的每个独特组来标识一个或多个相关的患者特性。

[0005] 所述机器学习模型可以被训练成使用一个或多个无监督聚类技术对所述一组患者进行分组。所述一个或多个无监督聚类技术可以包括二分k均值聚类技术。对所述一组患者进行分组可以包括执行多重对应分析以减少所述多个患者特性的维度。

[0006] 选择所述一组独特组可以包括:针对所述多个独特组中的每个独特组,确定由所述独特组表现出的每个患者特性的特征得分。选择所述一组独特组可以包括将所述多个独特组中的每个独特组的所述特征得分与特征得分阈值进行比较。选择一组独特组可以包括以下中的至少一个:确定所述多个独特组中的每个独特组的稳定性量度;确定所述多个独

特组中的每个独特组的纯度量度;以及确定所述一组患者的包括在所述多个独特组中的每个独特组中的患者数量。

[0007] 标识所述一个或多个相关的患者特性可以包括:针对所述一组独特组中的每个独特组,通过选择所述多个患者特性中的由所述独特组表现出的并且对应于所述组的主题的患者特性来生成多个潜在相关的患者特性。标识所述一个或多个相关的患者特性可以包括对所述多个潜在相关的患者特性中的所述潜在相关的患者特性中的每一个进行评级。对所述潜在相关的患者特性中的每一个进行评级可以包括:针对每个潜在相关的患者特性,基于所述潜在相关的患者特性和至少一个与所述药物相关联的参考适应证的同现频率分配等级值。所述同现可以是通过以下测量的:针对所述潜在相关的患者特性中的每一个,确定所述一组独特组的包括所述潜在相关的患者特性和所述至少一个参考适应证两者的比例。标识所述一个或多个相关的患者特性可以包括:针对所述潜在相关的患者特性中的每一个,确定临床可行性和商业可行性中的至少一个。

[0008] 所述一个或多个操作可以包括将所述一个或多个相关的患者特性中的至少一个相关的患者特性标识为用于再利用所述药物的目标适应证。

[0009] 这些和其他方面、特征和实现方式可以被表达为用于执行功能的方法、设备、系统、组件、程序产品、手段或步骤,以及其他方式。

[0010] 本公开文本的实现方式可以提供一个或多个以下优点。当与常规技术相比时,本说明书中描述的系统和方法可以通过例如减少用于处理大量数据的计算时间来提高计算效率,所述大量数据具有不同复杂性水平以标识潜在相关的适应证(在本说明书中有时被称为患者特性)。可以使用特定的机器学习技术来发现可能无法通过常规技术标识的新的适应证。当与常规技术相比时,本说明书中描述的系统和方法可以减少对人类输入和技能的依赖。

[0011] 从包括权利要求在内的以下描述中,这些和其他方面、特征和实现方式将变得显而易见。

附图说明

[0012] 图1是图示了用于药物再利用的数据处理系统的例子的图。

[0013] 图2是图示了用于再利用药物的示例方法的流程图。

[0014] 图3是图示了使用本说明书中描述的系统和方法进行的实验的图。

[0015] 图4是图示了由使用本说明书中描述的系统和方法产生的实验结果的图。

[0016] 图5是用于提供与本公开文本中描述的算法、方法、功能、过程、流程和程序相关联的计算功能的示例计算机系统的框图。

具体实施方式

[0017] 药物再利用可以用于发现临床批准的药物的新的临床适应证(例如,使用药物的原因)。随着探索出数百种新的适应证,关键意见领袖(KOL)可能无法解决与这些新的适应证相对应的高复杂性水平。处理大量数据和使用先进分析学可以支持以KOL为中心的方法。在相关临床问题的指导下,强大的先进分析学技术可以挖掘出隐藏在大量数据中的临床相关信息,所述临床相关信息然后可以帮助临床决策。

[0018] 计算药物再利用方法可以使用相似度量度(化学相似度、分子活性相似度、基因表达相似度或副作用相似度)、分子对接或共享的分子病理学来检测新的药物-疾病关系。药物再利用方法可以被分类为基于网络的文本挖掘(文献搜索)和语义方法。

[0019] 基于网络的方法可以涉及通过将比如药物、蛋白质、基因和疾病等多个数据来源相结合来创建集成网络。例如,联系图(C-Map)方法可以利用转录组,通过利用基因表达谱分析来将生物学、化学和临床病状联系起来,以促进发现新的疾病-基因-药物连接。基于网络的聚类技术(聚类)可以用于发现生物模块。这种方法是受生物网络的同一模块中的生物实体(疾病、药物、蛋白质等)通常共有相似的特性这一事实的启发。聚类可以涉及使用网络拓扑结构来发现药物-疾病关系、药物-药物关系、疾病-疾病关系或药物-靶标关系。实现聚类方法可能会涉及一些困难。例如,药物和疾病的边缘连接可能取决于收集到的药物-疾病关联,这些关联是不完整的,这可能需要整合多个数据库以提高预测的准确度。另外,并入提供关于药物副作用的信息的不同数据来源可以允许收集潜在的安全信号。此外,可能难以区分负关联和正关联,而且没有现成的“黄金”标准方法来测试生物模块之间的关联。

[0020] 基于文本挖掘的方法可以使用关键词同现和语义推断来进行新的药物-疾病关联。一些方法可以基于Swanson的类比推理方法(ABC模型),这种方法可以假设,如果“B”是疾病“C”的特性之一,并且物质“A”会影响“B”,那么可以通过B连接推导出“A:C”的隐含联系。然而,由于语言的模糊性质、对生物医学关系的有限覆盖以及文本挖掘技术的有限准确度,因此单独的基于文献挖掘的方法可能是有限的。

[0021] 本说明书中描述的数据处理系统和方法的实现方式可以通过使用用于药物再利用的真实世界数据驱动方案来缓解先前提到的缺点,以便标识适应证。在一些实现方式中,本说明书中描述的数据处理系统和方法将分析学和真实世界数据与KOL临床输出相结合。在一些实现方式中,本说明书中描述的数据处理系统和方法以无监督方式使用真实世界数据和分析学。在一些实现方式中,使用机器学习技术将患者聚类,并对跨多个簇出现的病状组进行标识。在一些实现方式中,所标识的病状组潜在地对应于常见生物途径。本说明书中描述的数据处理系统和方法可以用于标识先前使用常规技术没有标识出的潜在的适应证。

[0022] 在以下描述中,出于解释的目的,阐述了许多具体细节以便提供对本公开文本的透彻理解。然而,显而易见的是,本公开文本可以在没有这些具体细节的情况下实施。在其他实例中,众所周知的结构和装置以框图形式示出,以避免不必要地模糊本公开文本。

[0023] 在附图中,为了便于描述,示出了示意性元素的特定排列或排序,例如表示装置、模块、指令块和数据元素的排列或排序。然而,本领域的技术人员应该理解,附图中示意性元素的特定顺序或排列并不意味着需要特定的处理顺序或过程分离。进一步,在附图中包括示意性元素并不意味着在所有实现方式中都需要这样的元素,或者由这样的元素表示的特征在一些实现方式中不可以包括在其他元素中或者不可以与其他元素组合。

[0024] 进一步,在附图中,连接元素,例如实线或虚线或箭头,用于说明两个或多个其他示意性元素之间的连接、关系或关联,缺少任何这样的连接元素并不程序意味着不存在连接、关系或关联。换句话说,元素之间的一些连接、关系或关联没有在附图中示出,以免混淆本公开文本。此外,为了便于图示,单个连接元素用于表示元素之间的多个连接、关系或关联。例如,在连接元素代表信号、数据或指令的通信的情况下,本领域技术人员应该理解,这种元素代表一个或多个信号路径(例如,总线),如可能需要的,以影响通信。

[0025] 现在将详细参考实现方式，其例子在附图中图示。在以下详细描述中，阐述了许多具体细节，以便提供对各种描述的实现方式的透彻理解。然而，对于本领域普通技术人员来说显而易见的是，可以在没有这些具体细节的情况下实施所描述的各种实现方式。在其他实例中，没有详细描述众所周知的方法、程序、组件、电路和网络，以免不必要地模糊实现方式的各个方面。

[0026] 下文描述了几个特征，每个特征可以彼此独立使用或者与其他特征的任意组合一起使用。然而，任何单个的特征可能不能解决上文讨论的任何问题，或者可能只解决上文讨论的问题之一。本说明书中描述的任何特征都可能不能完全解决上面讨论的一些问题。尽管提供了标题，与特定标题相关但未在具有该标题的部分中找到的数据也可以在本说明书的其他地方找到。

示例数据处理系统和方法

[0027] 图1示出了数据处理系统100的例子。在一些实现方式中，数据处理系统100被配置成处理可以表示多个患者的医疗记录的数据以标识药物(用于再利用的药物)的新的适应证。系统100包括计算机处理器110。计算机处理器110包括计算机可读存储器111和计算机可读指令112。系统100还包括机器学习系统150。机器学习系统150包括机器学习模型120。机器学习模型120可以与计算机处理器110分离或集成。

[0028] 计算机可读介质111(或计算机可读存储器)可以包括适合于本地技术环境的任何数据存储技术类型，包括但不限于基于半导体的存储器装置、磁存储器装置和系统、光存储器装置和系统、固定存储器、可移动存储器、盘存储器、闪存、动态随机存取存储器(DRAM)、静态随机存取存储器(SRAM)、电可擦除可编程只读存储器(EEPROM)等。在一些实现方式中，计算机可读介质111包括具有可执行指令的代码段。

[0029] 在一些实现方式中，计算机处理器110包括通用处理器。在一些实现方式中，计算机处理器110包括中央处理单元(CPU)。在一些实现方式中，计算机处理器110包括至少一个专用集成电路(ASIC)。计算机处理器110还可以包括通用可编程微处理器、图形处理单元、专用可编程微处理器、数字信号处理器(DSP)、可编程逻辑阵列(PLA)、现场可编程门阵列(FPGA)、专用电子电路等、或它们的组合。计算机处理器110被配置成执行诸如计算机可执行指令112的程序代码，并且被配置成执行包括机器学习模型120的可执行逻辑。

[0030] 计算机处理器110被配置成接收表示多个患者的医疗记录的数据。例如，计算机处理器110可以从包括可通过关键标识符(ID)标识的大约9400万个患者(或更多个患者)的电子医疗记录(EMR)数据的数据库接收数据，所述关键标识符允许跨不同的数据表对患者进行匹配。在一些实现方式中，数据指示诊断、实验室测试、手术、药物治疗、患者事件、保险、生物标志物、测量结果、临床状态、生活方式参数、微生物学、处方等。在一些实现方式中，数据包括自然语言处理驱动数据。数据可以通过各种技术中的任何技术被接收，例如无线通信、光纤通信、USB、CD-ROM等。

[0031] 机器学习系统150能够应用机器学习技术来训练机器学习模型120。作为机器学习模型120的训练的一部分，机器学习系统150可以通过识别已经被确定为具有所述属性的输入数据项的正训练集来形成输入数据的训练集，并且在一些实现方式中，可以形成缺少所述属性的输入数据项的负训练集。

[0032] 机器学习系统150从训练集的输入数据中提取特征值，这些特征是被认为与输入

数据项是否具有相关属性潜在相关的变量。输入数据的特征的有序列表在这里被称为输入数据的特征向量。在一些实现方式中，机器学习系统150应用降维来将输入数据的特征向量中的数据量减少到更小、更具代表性的数据集。例如，机器学习系统150可以应用多重对应分析(MCA)、线性判别分析(LDA)、主成分分析(PCA)等。

[0033] 在一些实现方式中，机器学习系统150使用无监督机器学习来训练机器学习模型120。通常，无监督机器学习在不参考已知或标记的结果的情况下使用输入向量根据数据集进行推断。在一些实现方式中，机器学习系统150可以执行聚类，以将数据点划分为多个组，使得同一组中的数据点与同一组中的其他数据点更相似并且与其他组中的数据点不相似。在一些实现方式中，聚类包括执行K均值聚类，在这种聚类中，数据点的一个级别的非嵌入式分区是通过对数据集进行迭代分区创建的。也就是说，如果K是期望的簇数量，那么在每次迭代中，数据集被分区为K个不相交的簇。处理可以继续，直到指定的聚类准则功能值得到优化为止。在一些实现方式中，机器学习系统150被配置成执行二分K均值聚类。二分k均值聚类通常涉及在每个二分步骤处(例如，通过使用k均值)将一个簇分成两个子簇，直到获得k个簇为止。当与K均值聚类相比时，二分K均值聚类可能更有益，因为二分K均值聚类可以在K是相对较大的值时减少计算时间、可以产生大小相似的簇并且可以产生熵较小的簇。

[0034] 计算机处理器110被配置成执行计算机可执行指令112以执行一个或多个操作。在一些实现方式中，一个或多个操作包括接收表示多个患者的医疗记录的数据。例如，计算机处理器110可以从包括可通过关键标识符(ID)标识的大约9400万个患者(或更多个患者)的电子医疗记录(EMR)的数据库接收数据，所述关键标识符允许跨不同的数据表对患者进行匹配。在一些实现方式中，数据指示诊断、实验室测试、手术、药物治疗、患者事件、保险、生物标志物、测量结果、临床状态、生活方式参数、微生物学、处方等。在一些实现方式中，数据包括自然语言处理驱动数据。数据可以通过各种技术中的任何技术被接收，例如无线通信、光纤通信、USB、CD-ROM等。

[0035] 在一些实现方式中，一个或多个操作包括基于医疗记录选择一组患者。选择一组患者包括确定至少一个与用于再利用的药物相关联的目标信号传导途径。例如，如果用于再利用的药物是度普利尤单抗(Dupilumab)，那么计算机处理器110可以确定所述药物基于所述药物的已知功能调节白介素4(IL-4)和白介素13(IL-13)信号传导途径。选择一组患者还包括基于对应于与目标信号传导途径联系的诊断的一个或多个因素确定一个或多个指标。例如，可以使用诸如途径机制、相关临床病状、治疗类似物、数据和流行病学以及药品生命周期管理一致性等因素来搜索包括医疗数据库和医疗证据软件的来源，以标识与所确定的信号传导途径联系的疾病。这些疾病可以基于与所确定的信号传导途径联系的强度进行归类。类别可以包括重点组、中等组和广泛组。例如，返回到IL-4/IL-13例子，重点疾病组可以包括与IL-4/IL-13对Th2途径的作用机制具有直接关系的疾病，中等疾病组可以包括与IL-4/IL-13对Th2途径的作用机制具有间接关系的疾病，并且广泛疾病组可以包括与更广泛的炎性反应相关联的疾病。从重点组转到广泛组可能增加在选择一组患者时要考虑的指标数量并且可能降低分子碰撞的可能性。因此，在一些实现方式中，仅使用重点组或使用重点组和中等组来选择一组患者。在一些实现方式中，仅选择具有至少一个与所确定的信号传导途径相关联的诊断、药物治疗、实验室测试和/或手术的患者纳入一组患者中。稍后参考表1提供因素和指标的详细例子。

[0036] 在一些实现方式中，一个或多个操作包括确定一组患者的多个患者特性(在本说明书中有时被称为特征)，其中，所述一组患者中的每个患者表现出所述多个患者特性中的至少一个患者特性。确定多个患者特性可以包括分析最初接收到的数据，以标识广泛的患者特性，以捕获接收到的数据的全部或大部分。例如，广泛的患者特性可以对应于诊断(例如，免疫病状、糖尿病)、处方(例如，免疫药物、其他药物分类)、手术(例如，人类白细胞抗原分型)和实验室结果(例如，IgE异常高/低)。在一些实现方式中，确定多个患者特性包括接收用户输入(例如，通过与计算机处理器110通信的用户接口)。例如，用户可以基于临床输入、人口统计资料、药物治疗、合并症、手术和对免疫学具有特异性的实验室测试数据输入患者特性。也可以添加定制的特性分类，以提高数据完整性、代表性并且收集更多关于疾病和药物应答的信息。在一些实现方式中，确定多个患者特性包括验证所述多个患者特性。验证可以包括通过计算具有每个特性族中的至少一个特性的所选患者的百分比(例如，具有处方记录的患者的百分比)并且将此百分比与最初接收到的数据的具有每个特性族中的至少一个特性的患者的百分比进行比较来确定最初接收到的数据的患者特性是否正确地被映射到所选的一组患者。两个数字的值较接近指示映射已经正确地完成。验证可以包括通过标识同时包括在最初接收到的数据和所选的一组患者中的多个患者以验证最初接收到的数据的患者与所选的一组患者之间相同的患者特性映射来确定患者特性是否已经被映射到正确的患者。

[0037] 在一些实现方式中，一个或多个操作包括根据多个患者特性(例如，如与所确定的信号传导途径相关的特征所定义的)对一组患者进行分组，以生成多个独特组，其中，所述独特组中的每个独特组包括所述一组患者中的至少一个患者。例如，一个或多个计算机处理器110可以执行机器学习模型120以执行聚类技术，例如上文所描述的二分k均值聚类技术。聚类可以产生多个患者簇(例如，多个独特的患者组)，其中，在患者的对应的患者特性方面，一个簇中的患者彼此相似的程度高于所述簇中的患者与其他簇中的患者相似的程度。在一些实现方式中，所生成的簇可以显示出患者特性之间的相互关系，即使这些患者特性并不存在于同一个患者中。临床输入可以在聚类过程的各个阶段中被接收并使用，以确保所产生的簇的临床相关性。例如，疾病专家的临床输入可以促进临床相关队列的创建、临床相关特征的纳入和分组以及对簇进行验证和评估。如果患者特性出现的频率高于其在一般群体中出现的频率(例如，就所选的一组患者总体而言)，那么所述患者特性可以被标识为在簇中是独特的。

[0038] 在一些实现方式中，使用多重对应分析(MCA)以减少患者特性的维度。二分K均值可以促进患者以足够“紧密”但稳定的簇适当且有效地分离，并且允许使用大量的表现出免疫相关性的簇对患者特性进行评分，稍后会对此进行更详细的解释。所产生的簇可以被呈现(例如，通过用户接口)给用户(例如，临床专家)以进行验证和评估。这可以降低簇的不可解释性的风险并且确保不同的簇之间不存在重叠的特征。

[0039] 在一些实现方式中，一个或多个操作包括基于一个或多个组选择标准选择多个独特组中的一组独特组。在一些实现方式中，选择一组独特组包括对组进行评级并且选择多个评级最高的组(例如，评级前60的组)。这些组可以基于免疫学富集度、稳定性、纯度和大小进行评级。在一些实现方式中，计算每个患者特性的一个或多个量度(在本说明书中有时被称为特征得分)以对簇进行评级。一个或多个量度可以包括例如独特性(在本说明书中有

时被称为“提升得分”)、簇内呈现出患者特性的患者数量以及免疫学得分。独特性得分测量患者特性在某个簇内相比于群体的其余部分的独特程度(例如,如果男性占群体的50%,占簇的75%,那么“提升得分”可以等于1.5)。在一些实现方式中,只有提升得分超过阈值提升得分(例如,1)并且在超过阈值患者百分比(例如,10%)的患者百分比中出现的患者特性被考虑用于定义簇并且对应于所述簇的主题。被考虑用于定义簇的患者特性可以在本说明书中被称为潜在相关的患者特性。然后可以给予患者特性(例如,被考虑用于定义簇的患者特性或所有患者特性)免疫学得分,所述免疫学得分根据所述患者特性的类型(例如,疾病、药物、实验室测试、手术等)和免疫学相关性对所述患者特性进行评分。然后可以对每个簇内的患者特性得分进行合计(例如,求和)和归一化。然后,符合阈值簇得分(例如,50%)的簇可以被认为是免疫学特异性的。

[0040] 选择一组独特组可以包括对每个簇内的稳定性、纯度和患者数量中的一个或多个进行评估。稳定性可以使用以下方法中的一种或多种方法进行评估:(1)在不同大小的数据上重现簇;(2)改变簇的初始化种子;(3)改变所产生的簇的数量;以及(4)应用训练-测试方法。对于训练集中的每个簇,稳定性可以被定义为在测试集中也被分组在一起的患者的最大比例。纯度可以通过簇内的患者的MCA分量的簇内方差来测量,这可以产生同质且密集的簇。在一些实现方式中,如果簇超过阈值稳定性百分比(例如,50%)并且超过阈值纯度百分比(例如,所述簇的纯度在所有簇的最高的20%中),那么选择所述簇。

[0041] 在一些实现方式中,一个或多个操作包括通过分析一组独特组中的每个独特组来标识一个或多个相关的患者特性(例如,适应证)。标识一个或多个相关的患者特性可以包括对每个所选簇所呈现出的患者特性(例如,所有患者特性或被考虑用于定义簇的患者特性)进行评级。评级可以基于与用于再利用的药物的多个既定(参考)特性(用作参考)中的每一个的同现频率(例如,如果药物是度普利尤单抗,那么参考特性可以包括哮喘、特应性皮炎、IgE过敏和综合免疫学得分)。同现可以通过计算含有患者特性和参考两者的患者加权簇的比例来测量。在一些实现方式中,由学科专家判断为与核心簇主题(例如,如通过用户接口接收到的用户输入所指示的)相关的一个或多个患者特性也可以被考虑用于进行评价,而不管出现这些特征的患者数量(可能<10%)如何。

[0042] 标识一个或多个患者特性可以包括评估患者特性的临床和商业可行性。例如,显示出独特的临床诊断的患者特性可以被标识。商业评估可以基于指示预测销售额和竞争对手资产的数据可用、与作为目标的信号途径的确定的联系(无论是否在出版物中找到)、患者特性的世界范围的流行率、以及患者特性的伤残调整生命年(DALY)(例如,每100,000生命年)。在一些实现方式中,一个或多个患者特性中的至少一个患者特性可以被标识为用于再利用药物的目标适应证。因此,在一些实现方式中,一个或多个操作通常输出用于再利用的药物的一个或多个新的适应证。

[0043] 虽然本说明书通常将患者描述为人类患者,但是实现方式并不局限于此。例如,患者可以指非人动物、植物或人类复制系统。

[0044] 虽然为了说明的目的,本说明书通常描述接收对应于大约9400万个患者的数据,但应当理解,数据可以对应于更少或更多的患者。

[0045] 虽然为了说明的目的,本说明书将用于再利用的药物描述为度普利尤单抗,但应当理解,用于再利用的药物可以是任何药物。

[0046] 图2是图示了用于再利用药物的示例方法200的流程图。方法200可以由之前参考图1描述的数据处理系统100执行。方法200包括:接收表示医疗记录的数据(框210);选择一组患者(框220);确定多个患者特性(框230);对所述一组患者进行分组,以生成多个独特组(框240);选择一组独特组(框250);以及标识一个或多个相关的患者特性(框260)。

[0047] 在框210处,接收表示多个患者的医疗记录的数据。例如,可以从包括可通过关键ID标识的大约9400万个患者(或更多个患者)的EMR的数据库接收数据,所述关键ID允许跨不同的数据表对患者进行匹配。在一些实现方式中,数据指示诊断、实验室测试、手术、药物治疗、患者事件、保险、生物标志物、测量结果、临床状态、生活方式参数、微生物学、处方、医学影像等。在一些实现方式中,数据包括自然语言处理驱动数据。数据可以通过各种技术中的任何技术被接收,例如无线通信、光纤通信、USB、CD-ROM等。

[0048] 在框220处,确定至少一个与用于再利用的药物相关联的目标信号传导途径。例如,如果药物是度普利尤单抗,那么可以确定所述药物基于所述药物的已知功能调节白介素4(IL-4)和白介素13(IL-13)信号传导途径。在一些实现方式中,基于对应于与目标信号传导途径联系的诊断的一个或多个因素确定一个或多个指标。例如,可以使用诸如途径机制、相关临床病状、治疗类似物、数据和流行病学以及药品生命周期管理一致性等因素来搜索包括医疗数据库和医疗证据软件的来源,以标识与所确定的信号传导途径联系的疾病。这些疾病可以基于与所确定的信号传导途径联系的强度进行归类。类别可以包括重点组、中等组和广泛组。例如,返回到IL-4/IL-13例子,重点疾病组可以包括与IL-4/IL-13对Th2途径的作用机制具有直接关系的疾病,中等镜头数据集疾病组可以包括与IL-4/IL-13对Th2途径的作用机制具有间接关系的疾病,并且广泛疾病组可以包括与更广泛的炎性反应相关联的疾病。从重点组转到广泛组可能增加在选择一组患者时要考虑的指标数量并且可能降低分子碰撞的可能性。因此,在一些实现方式中,仅使用重点组或使用重点组和中等组来选择一组患者。在一些实现方式中,仅选择具有至少一个与所确定的信号传导途径相关联的诊断、药物治疗、实验室测试和/或手术的患者纳入一组患者中。

[0049] 在框230处,确定一组患者的多个患者特性,其中,所述一组患者中的每个患者表现出所述多个患者特性中的至少一个患者特性。确定多个患者特性可以包括分析最初接收到的数据,以标识广泛的患者特性,以捕获接收到的数据的全部或大部分。例如,广泛的患者特性可以对应于诊断(例如,免疫病状、糖尿病)、处方(例如,免疫药物、其他药物分类)、手术(例如,人类白细胞抗原分型)和实验室结果(例如,IgE异常高/低)。在一些实现方式中,确定多个患者特性包括接收用户输入(例如,通过用户接口)。例如,用户可以基于临床输入和人口统计资料、药物治疗、合并症、手术和对免疫学具有特异性的实验室测试数据输入患者特性。也可以添加定制的特性分类,以提高数据完整性、代表性并且收集更多关于疾病和药物应答的信息。在一些实现方式中,确定多个患者特性包括验证所述多个患者特性。验证可以包括通过计算具有每个特性族中的至少一个特性的所选患者的百分比(例如,具有处方记录的患者的百分比)并且将此百分比与最初接收到的数据的具有每个特性族中的至少一个特性的患者的百分比进行比较来确定最初接收到的数据的患者特性是否正确地被映射到所选的一组患者。两个数字的值较接近指示映射已经正确地完成。验证可以包括通过标识同时包括在最初接收到的数据和所选的一组患者中的多个患者以验证最初接收到的数据的患者与所选的一组患者之间相同的患者特性映射来确定患者特性是否已经被

映射到正确的患者。

[0050] 在框240处，根据多个患者特性(例如，如与所确定的信号传导途径相关的特征所定义的)对一组患者进行分组，以生成多个独特组，其中，所述独特组中的每个独特组包括所述一组患者中的至少一个患者。例如，可以使用多个患者特性对一组患者执行聚类技术，例如上文所描述的二分k均值聚类技术。聚类可以产生多个患者簇(例如，多个独特的患者组)，其中，在患者的对应的患者特性方面，一个簇中的患者彼此相似的程度高于所述簇中的患者与其他簇中的患者相似的程度。在一些实现方式中，所生成的簇可以显示出患者特性之间的相互关系，即使这些患者特性并不存在于同一个患者中。临床输入可以在聚类过程的各个阶段中被接收并使用，以确保所产生的簇的临床相关性。例如，疾病专家的临床输入可以促进临床相关队列的创建、临床相关特征的纳入和分组以及对簇进行验证和评估。如果患者特性出现的频率高于其在一般群体中出现的频率(例如，就所选的一组患者总体而言)，那么所述患者特性可以被标识为在簇中是独特的。

[0051] 在一些实现方式中，使用多重对应分析(MCA)以减少患者特性的维度。二分K均值可以促进患者以足够“紧密”但稳定的簇适当且有效地分离，并且允许使用大量的表现出免疫相关性的簇对患者特性进行评分，稍后会对此进行更详细的解释。所产生的簇可以被呈现(例如，通过用户接口)给用户(例如，临床专家)以进行验证和评估。这可以降低簇的不可解释性的风险并且确保不同的簇之间不存在重叠的特征。

[0052] 在框250处，基于一个或多个组选择标准选择多个独特组中的一组独特组。在一些实现方式中，选择一组独特组包括对组进行评级并且选择多个评级最高的组(例如，评级前60的组)。这些组可以基于免疫学富集度、稳定性、纯度和大小进行评级。在一些实现方式中，计算每个患者特性的一个或多个量度以对簇进行评级。一个或多个量度可以包括例如独特性(在本说明书中有时被称为“提升得分”)、簇内呈现出患者特性的患者数量以及免疫学得分。独特性得分测量患者特性在某个簇内相比于群体的其余部分的独特程度(例如，如果男性占群体的50％，占簇的75％，那么“提升得分”可以等于1.5)。在一些实现方式中，只有提升得分超过阈值提升得分(例如，1)并且在超过阈值患者百分比(例如，10％)的患者百分比中出现的患者特性被考虑用于定义簇并且对应于所述簇的主题。被考虑用于定义簇的患者特性可以在本说明书中被称为潜在相关的患者特性。然后可以给予患者特性(例如，被考虑用于定义簇的患者特性或所有患者特性)免疫学得分，所述免疫学得分根据所述患者特性的类型(例如，疾病、药物、实验室测试、手术等)和免疫学相关性对所述患者特性进行评分。然后可以对每个簇内的患者特性得分进行合计(例如，求和)和归一化。然后，符合阈值簇得分(例如，50％)的簇可以被认为是免疫学特异性的。

[0053] 选择一组独特组可以包括对每个簇内的稳定性、纯度和患者数量中的一个或多个进行评估。稳定性可以使用以下方法中的一种或多种方法进行评估：(1)在不同大小的数据上重现簇；(2)改变簇的初始化种子；(3)改变所产生的簇的数量；以及(4)应用训练-测试方法。对于训练集中的每个簇，稳定性可以被定义为在测试集中也被分组在一起的患者的最大比例。纯度可以通过簇内的患者的MCA分量的簇内方差来测量，这可以产生同质且密集的簇。在一些实现方式中，如果簇超过阈值稳定性百分比(例如，50％)并且超过阈值纯度百分比(例如，所述簇的纯度在所有簇的最高的20％中)，那么选择所述簇。

[0054] 在框260处，通过分析一组独特组中的每个独特组标识一个或多个相关的患者特

性。标识一个或多个相关的患者特性可以包括对每个所选簇所呈现出的患者特性(例如,所有患者特性或被考虑用于定义簇的患者特性)进行评级。评级可以基于与用于再利用的药物的每一多个既定(参考)特性(用作参考)的同现频率(例如,如果药物是度普利尤单抗,那么参考特性可以包括哮喘、特应性皮炎、IgE过敏和综合免疫学得分)。同现可以通过计算含有患者特性和参考两者的患者加权簇的比例来测量。在一些实现方式中,由学科专家判断为与核心簇主题(例如,如通过用户接口接收到的用户输入所指示的)相关的一个或多个患者特性也可以被考虑用于进行评价,而不管出现这些特征的患者数量(可能<10%)如何。

[0055] 标识一个或多个患者特性可以包括评估患者特性的临床和商业可行性。例如,显示出独特的临床诊断的患者特性可以被标识。商业评估可以基于指示预测销售额和竞争对手资产的数据可用、与作为目标的信号途径的确定的联系(无论是否在出版物中找到)、患者特性的世界范围的流行率、以及患者特性的伤残调整生命年(DALY)(例如,每100,000生命年)。在一些实现方式中,一个或多个患者特性中的至少一个患者特性可以被标识为用于再利用药物的目标适应证。

实验结果

[0056] 图3是图示了使用本说明书中描述的系统和方法进行的实验的图。进行实验以验证用于度普利尤单抗(其是抗IL4/IL13药物)的药物再利用的RWD驱动方案,以便标识药物的新型适应证。实验的一个目标是减少药物开发成本和上市时间,同时使损耗和风险最小化。通过KOL专业知识、商业评估以及与真实世界数据相结合的分析学,利用科学和临床能力,采用了混合方法。

[0057] 数据来源:使用了2014年至2018年的Optum Humedica数据集。数据库含有可通过关键标识符标识的9400个患者的电子医疗记录,所述关键标识符允许跨不同的数据表对患者进行匹配。数据库收集了关于EMR数据的信息,比如诊断、实验室测试、手术、药物治疗、患者事件、保险、生物标志物、测量结果、临床状态和生活方式参数、微生物学和处方。自然语言处理(NLP)驱动表由于有限的数据覆盖和临床相关性而没有包括在内。此外,排除了不完整或含有不相关信息的数据表。总共包括5个数据表,将数据来源减少至4000万个患者。

[0058] 患者选择:患者选择的指标基于与潜在的免疫学途径相关的临床框架,如表1所示。

因素	重点镜头数据集（lens）-引起疾病的Th2功能障碍	中等镜头数据集-与疾病相关联的Th2应答	广泛镜头数据集-貌似与疾病相关联的Th2应答
途径机制	嗜酸性粒细胞增多症 IL-4途径 IL-5途径 IL-13途径	1型超敏反应 4型超敏反应 瘙痒症	相关自身免疫病状 其他自身免疫和/或炎性途径 末梢器官炎性病状（例如，心脏炎性病状、呼吸系统炎性病状、肾脏炎性病状） 肿瘤学途径
相关临床病状	嗜酸性食管炎 嗜酸性肉芽肿性多血管炎(查格-施特劳斯综合征（Churg-Strauss Syndrome）） 嗜酸性病状	过敏性反应 过敏性结膜炎 荨麻疹	甲状腺炎 胰腺炎 淀粉样变性 基底细胞癌
治疗类似物	IL-4抑制剂 IL-5抑制剂 IL-13抑制剂	下游途径调节剂 JAK抑制剂 STAT抑制剂 GATA抑制剂 IL-25 上游途径调节剂 IL-2抑制剂	其他途径，其确定白细胞介素目标： IL-1、IL-6、IL-12、IL-21、IL-23 下游白介素目标： IL-9、IL-17、IL-22 其他潜在的促炎性目标： IFN、TGFβ、TNFα
数据和流行病学	嗜酸性病状	1型超敏反应病状	自身免疫病状 炎性病状 肿瘤学疾病
LCM策略一致性	嗜酸性食管炎 花生和草类过敏的辅助疗法	慢性鼻窦炎 慢性自发性荨麻疹 特应性角膜结膜炎	

表1:患者选择考虑到的用于标识适应证的5个因素

仅选择了曾经患有IL4/13途径(即,信号传导途径)相关诊断的具有至少一种诊断、药物治疗、实验室测试和手术的成年患者(年龄≥18岁)。使用免疫学病状和数据完整性的这些标准,所产生的队列由1700万名唯一患者组成。

[0059] 对指标进行标识考虑了五个因素(如表2所示)。这些因素是通过Doctor Evidence引擎数据库(一家医疗证据软件和服务公司,由多个平台(DOC Library、DOC Data、Doctor Evidence、DOC Label、DOC Search)组成并且包括PubMed、ClinicalTrials.gov、WHO等))上

的来源搜索到的。然后，基于Th2应答根据三个镜头数据集对这些因素内的信息进行分类：重点、中等和广泛。例如，基于疾病与IL4/IL13对Th2途径的作用机制的直接和间接关系，将所述疾病分别分配到重点镜头数据集或中等镜头数据集，如果所述疾病与较广泛的炎性反应相关联，则将所述疾病分配到广泛镜头数据集。从重点镜头数据集转到广泛镜头数据集可能增加要考虑的指标范围并且可能降低分子碰撞的可能性。广泛镜头数据集中的指标最终被排除在分析之外，因为它们的机理联系不够特异性，无法满足标识具有相似特性的患者群体的标准。因此，实验中仅包括重点镜头数据集和中等镜头数据集的指标。最终列表包括208个指标，用于跨17个广泛系统进行分析。

[0060] 特征选择：选择广泛特征(患者特性)以捕获Optum数据集中的可用信息；然后，对临床专家所选择的特征进行优先级排序和验证，以确保包括所有基本变量，并且变量值具有临床意义。在适当的时候，保留某些特征，重新创建其他特征(某些人口统计资料)(如稍后参考图4所描述的V1所示出的)。基于临床输入和人口统计资料、药物治疗、合并症、手术和对免疫学具有特异性的实验室测试数据添加新的特征。创建定制的特征分类并将所述定制的特征分类迭代地添加到分析中(如稍后参考图4所描述的V2和V3所示出的)，以提高数据完整性、代表性并且收集更多关于疾病和药物应答的严重程度的信息。

[0061] 使用稳健方法，以确保最终数据库中的特征的完整性。进行两个验证步骤(基于跨Optum数据库和所生成的表进行的患者和特征映射)，以验证特征正确地生成。首先，为了验证Optum中的患者特征是否正确地被映射到数据表，计算具有至少一个特征族的患者百分比，并且确定此数字在Optum Humedica和数据表中是否相同。其次，为了验证特征是否被映射到正确的患者，从原始Optum Humedica数据到所生成的数据集追踪十个随机患者，以确保两个数据集中的特征的映射相同。在特征验证已经证实了正确的映射之后，用所包括的2700个特征对1700万患者运行算法。

[0062] 聚类：使用聚类技术将共有如与IL4/13途径相关的特征所定义的相似特性的患者分组在一起。聚类基于患者的特征寻找所述患者之间的相似性。所生成的簇使得发现了病状之间的相互关系，即使这些病状并不存在于同一个患者中。在过程的各个阶段中植入临床输入，以确保结果的临床相关性。因此，疾病专家的临床输入有助于临床相关队列的创建，有助于临床相关特征的纳入和分组，并且最终有助于簇验证和评估。如果特征出现的频率高于其在一般群体中出现的频率，那么将所述特征标识为在簇中是独特的。

[0063] 使用多重对应分析(MCA)以减少特征的维度。然后，利用二分K均值将数据分成500个簇，以提供患者以足够“紧密”但稳定的簇进行的适当且有效的分离并且允许使用大量的表现出免疫相关性的簇进行适应证评分。由临床专家对通过过程标识出的簇进行验证和评估。此步骤促进降低簇的不可解释性的风险并且确保不同的簇之间不存在重叠的特征。聚类方法运行了表示2700个特征(例如，MCA分量)和1700万个患者的数据。在算法结束时产生的簇数量为500。

[0064] 标识新的适应证(即，相关的患者特性)：进行进一步的评估、临床和商业判断，以获得较短的优先级信号列表并且基于簇输出跨簇标识最具临床相关性的适应证。使用四个方法步骤。第一个方法步骤基于免疫学富集度、稳定性、纯度和大小选择在500个簇之中评级前60的簇。计算每个簇中所包括的特征的三个量度以确定选择：独特性、每个簇内呈现出特征的患者数量以及免疫学得分。独特性(也被称为“提升得分”)测量特征在某个簇内相比

于群体的其余部分的独特程度(例如,如果男性占群体的50%,占簇的75%,那么提升得分等于1.5)。只有提升得分>1(这意味着相较于较广泛的群体而言特征在簇中出现的比率高于预期)并且在≥10%的患者中出现的特征被考虑用于定义(和命名)簇以及确立簇的主题。另外,根据每个所选特征的类型(疾病、药物、实验室测试或手术)和免疫学相关性给予所述每个所选特征一个得分。对每个簇内的特征得分进行相加和归一化。如果簇满足50%的得分这一预定义阈值,那么所述簇可以被认为是免疫学特异性的。作为第二个步骤,根据稳定性、纯度、患者数量选择簇。稳定性可以使用以下四种方法进行评估:1.在不同大小的数据上重现簇;2.改变簇的初始化种子;3.改变所产生的簇的数量;以及4.应用训练-测试方法。对于训练集中的每个簇,稳定性被定义为在测试集中也被分组在一起的患者的最大比例。纯度可以通过簇内的患者的MCA分量的簇内方差来测量,从而产生同质且密集的簇。如果簇的稳定性大于50%并且其纯度在最高的20%中,那么将所述簇纳入分析中。另外,由学科专家判断为与核心簇主题相关的所有适应证被考虑用于进行评价,而不管出现这些特征的患者数量(可能<10%)如何。然后,在第三个步骤中,基于与度普利尤单抗的四个既定适应证(用作参考)(哮喘、特应性皮炎、IgE过敏和综合免疫学得分)中的每个既定适应证的同现频率对这些新的适应证进行评级。同现是通过计算含有适应证和参考两者的患者加权簇的比例来测量的。在最后一个步骤中,通过临床和商业可行性对适应证最终列表进行进一步表征。临床评估保留了显示出独特的临床诊断的适应证。基于以IL4/IL13用作参考进行的评级,删除了前30中没有出现的病状,因为这些病状看起来与IL4/IL13调节关系不大。只有在关于预测销售额和竞争对手资产的数据可用的情况下,才有可能仅对临床合理的适应证子集进行商业评估。另外,多个因素也被考虑用于进行商业评估:与IL4/13途径的联系(无论是否在文献中找到)、适应证的世界范围的流行率、以及适应证的伤残调整生命年(DALY)(每100,000生命年)。

[0065] 结果:图4是图示了由使用本说明书中描述的系统和方法产生的实验结果的图。对从Optum Humedica提取的1700万个患者的最终队列进行分析,以评估数据跨所选特征的完整性和代表性。在中等镜头数据集群体中,患者队列的构成中59%为女性,大部分为高加索人(77%),最后一次活动的平均年龄为53岁(SD=7岁),并且平均随访期为7.1年。患者最常见地呈现出作为免疫学病状的急性鼻窦炎(25.2%)、过敏性鼻炎、不明鼻炎(20.6%)及其他以及不明哮喘(19.4%)的ICD10代码。最常见的免疫学相关药物治疗是泼尼松(prednisone)(28.0%)、糠酸氟替卡松(fluticasone furoate)(22.2%)和甲基强的松龙(methylprednisolone)(13.3%),并且0.4%的患者接受了过敏原免疫疗法注射和β2糖蛋白抗体测定。测试了大部分患者的白细胞计数(70.8%)、绝对中性粒细胞计数(ANC)(64.3%)和绝对淋巴细胞计数(ALC)(63.8%)。聚类程序创建了500个簇,其中125个簇被分类为既富集免疫病状又稳定。在这125个簇中,110个簇还被分类为纯的。在这些簇中,保留了含有最大的每簇患者数量的60个簇,并对其进行了临床相关信号的分析。在验证过程之后,使用训练-测试方法,84%的簇被认为是高度稳定的,无论种子位置和数据表中的患者数量分别如何,90%和99%的前20个簇都被重现。基于簇中所包括的特征标识了六个簇主题,并且还在V2和V3迭代中部分地重现:多器官免疫影响、瘤形成、哮喘和其他超敏反应、肌肉骨骼功能障碍、心脏代谢谱系以及妇产科病状。其中,通过簇评估选出了250个适应证,并通过与每个参考的同现进行了评级。通过临床和商业可行性对约85个适应证的列表进行了

进一步表征:其中,约20个适应证不表现独特的临床诊断或IL4/13调节的临床原理不充分,其他适应证没有现成的商业评估信息。来自混合方法的适应证最终列表标识了大约90%的已经在生命周期管理中的适应证,以及大约60%的附加潜在新适应证。

[0066] 图5是根据本公开文本的一些实现方式的示例计算机系统600的框图,所述示例计算机系统用于提供与本公开中描述的算法、方法、功能、过程、流程和程序(例如先前参考图2描述的方法200)相关联的计算功能。图示的计算机602旨在包括任何计算装置,例如服务器、台式计算机、膝上型/笔记本计算机、无线数据端口、智能电话、个人数据助理(PDA)、平板计算装置或这些装置中的一个或多个处理器,包括物理实例、虚拟实例或两者。计算机602可以包括能够接受用户信息的输入装置,例如小键盘、键盘和触摸屏。此外,计算机602可以包括输出装置,该输出装置可以传送与计算机602的操作相关联的信息。该信息可以包括数字数据、视觉数据、音频信息或信息的组合。信息可以在图形用户接口(UI)(或GUI)中呈现。

[0067] 计算机602可以充当客户端、网络组件、服务器、数据库、持久性或用于执行本公开文本中描述的主题的计算机系统的组件。图示的计算机602与网络630可通信地联接。在一些实现方式中,计算机602的一个或多个组件可以被配置成在不同的环境中操作,包括基于云计算的环境、本地环境、全局环境以及环境的组合。

[0068] 在高级别上,计算机602是可操作来接收、传输、处理、存储和管理与所描述的主题相关联的数据和信息的电子计算装置。根据一些实现方式,计算机602还可以包括应用服务器、电子邮件服务器、网络服务器、缓存服务器、流数据服务器或服务器的组合,或者与之可通信地联接。

[0069] 计算机602可以通过网络630从客户端应用程序(例如,在另一台计算机602上执行的)接收请求。计算机602可以通过使用软件应用处理接收到的请求来响应接收到的请求。请求也可以从内部用户(例如,从命令控制台)、外部(或第三方)、自动化应用、实体、个人、系统和计算机发送到计算机602。

[0070] 计算机602的每个组件可以使用系统总线603进行通信。在一些实现方式中,计算机602的任何或所有组件,包括硬件或软件组件,可以通过系统总线603彼此接口或与接口604(或两者的组合)接口。接口可以使用应用编程接口(API)612、服务层613或API 612和服务层613的组合。API 612可以包括例程、数据结构和对象类的规范。API 612可以独立于计算机语言,也可以依赖于计算机语言。API 612可以指完整的接口、单个功能或一组API。

[0071] 服务层613可以向计算机602和可通信地联接到计算机602的其他组件(无论是否示出)提供软件服务。使用该服务层的所有服务消费者都可以访问计算机602的功能。诸如由服务层613提供的软件服务可以通过定义的接口提供可重用的、定义的功能。例如,接口可以是用JAVA、C++或以可扩展标记语言(XML)格式提供数据的语言编写的软件。虽然被示出为计算机602的集成组件,但是在替代实现方式中,API 612或服务层613可以是与计算机602的其他组件以及可通信地联接到计算机602的其他组件相关的独立组件。此外,在不脱离本公开文本的范围的情况下,API 612或服务层613的任何或所有部分可以被实现为另一软件模块、企业应用或硬件模块的子模块或子模块。

[0072] 计算机602包括接口604。尽管在图5中被示为单个接口604,但是根据计算机602和所描述的功能的特定需求、期望或特定实现方式,可以使用两个或更多个接口604。计算机

602可以使用接口604与分布式环境中连接到网络630(无论是否示出)的其他系统通信。通常,接口604可以包括或使用编码在可操作来与网络630通信的软件或硬件(或软件和硬件的组合)中的逻辑来实现。更具体地,接口604可以包括支持与通信相关联的一个或多个通信协议的软件。这样,网络630或接口的硬件可用于在所示计算机602内部和外部传送物理信号。

[0073] 计算机602包括处理器605。尽管在图5中被示为单个处理器605,但是根据计算机602和所描述的功能的特定需求、期望或特定实现方式,可以使用两个或更多个处理器605。通常,处理器605可以执行指令并且可以操纵数据来执行计算机602的操作,包括使用如本公开文本中描述的算法、方法、功能、过程、流程和程序的操作。

[0074] 计算机602还包括数据库606,该数据库可以保存计算机602和连接到网络630的其他组件(无论是否示出)的数据。例如,数据库606可以是存储器内的、传统的或存储与本公开文本一致的数据的数据库。在一些实现方式中,根据计算机602和所述功能的特定需求、期望或特定实现方式,数据库606可以是两种或更多种不同数据库类型的组合(例如,混合存储器内数据库和传统数据库)。尽管在图5中被示为单个数据库606,但是根据计算机602和所描述的功能的特定需求、期望或特定实现方式,可以使用两个或更多个数据库(相同类型、不同类型或类型的组合)。虽然数据库606被示为计算机602的内部组件,但是在替代实现方式中,数据库606可以在计算机602的外部。

[0075] 计算机602还包括存储器607,其可以保存计算机602或连接到网络630的组件的组合(无论是否示出)的数据。存储器607可以存储符合本公开文本的任何数据。在一些实现方式中,根据计算机602和所描述的功能的特定需求、期望或特定实现方式,存储器607可以是两种或多种不同类型的存储器的组合(例如,半导体和磁存储器的组合)。尽管在图5中被示为单个存储器607,但是根据计算机602和所描述的功能的特定需求、期望或特定实现方式,可以使用两个或更多个存储器607(相同、不同或类型的组合)。虽然存储器607被示为计算机602的内部组件,但是在替代实现方式中,存储器607可以在计算机602的外部。

[0076] 应用608可以是算法软件引擎,其根据计算机602和所描述的功能的特定需求、期望或特定实现方式来提供功能。例如,应用608可以充当一个或多个组件、模块或应用。此外,尽管被示为单个应用608,但是应用608可以被实现为计算机602上的多个应用608。另外,尽管图示为在计算机602内部,但是在替代实现方式中,应用608可以在计算机602外部。

[0077] 计算机602还可以包括电源614。电源614可以包括可充电或不可充电的电池,该电池可以被配置成用户可更换或用户不可更换。在一些实现方式中,电源614可以包括功率转换和管理电路,包括再充电、待机和功率管理功能。在一些实现方式中,电源614可以包括电源插头,以允许计算机602插入墙壁插座或电源,例如,给计算机602供电或给可充电电池再充电。

[0078] 可以有任意数量的计算机602与包含计算机602的计算机系统相关联或在计算机系统外部,每个计算机602通过网络630通信。此外,在不脱离本公开文本的范围的情况下,术语“客户端”、“用户”和其他适当的术语可以适当地互换使用。此外,本公开文本设想许多用户可以使用一台计算机602,并且一个用户可以使用多台计算机602。

[0079] 本说明书中描述的主题和功能操作的实现方式可以在数字电子电路中、在有形体现的计算机软件或固件中、在包括本说明书中公开的结构及其结构等同物的计算机硬件

中、或者在它们中的一个或多个的组合中实现。所描述主题的软件实现方式可以被实现为一个或多个计算机程序。每个计算机程序可以包括编码在有形的、非暂时性的、计算机可读的计算机存储介质上的计算机程序指令的一个或多个模块，用于由数据处理设备执行或控制数据处理设备的操作。替代性地或附加地，程序指令可以被编码在人工生成的传播信号中/上。例如，信号可以是机器生成的电、光或电磁信号，其被生成以编码信息，用于传输到合适的接收器设备，以由数据处理设备执行。计算机存储介质可以是机器可读存储装置、机器可读存储基底、随机或串行存取存储装置或计算机存储介质的组合。

[0080] 术语“数据处理设备”、“计算机”和“电子计算机装置”(或本领域普通技术人员理解的等同物)指的是数据处理硬件。例如，数据处理设备可以包括用于处理数据的所有种类的设备、装置和机器，包括例如可编程处理器、计算机或多个处理器或计算机。该设备还可以包括专用逻辑电路，包括例如中央处理器(CPU)、现场可编程门阵列(FPGA)或专用集成电路(ASIC)。在一些实现方式中，数据处理设备或专用逻辑电路(或数据处理设备或专用逻辑电路的组合)可以是基于硬件或软件的(或基于硬件和软件的组合)。该设备可以可选地包括为计算机程序创建执行环境的代码，例如，构成处理器固件、协议栈、数据库管理系统、操作系统或执行环境的组合的代码。本公开文本涵盖使用具有或不具有常规操作系统的数据处理设备，例如LINUX、UNIX、WINDOWS、MAC OS、ANDROID或IOS。

[0081] 计算机程序也可以被称为或描述为程序、软件、软件应用、模块、软件模块、脚本或代码，可以用任何形式的编程语言编写。编程语言可以包括例如编译语言、解释语言、声明语言或过程语言。程序可以以任何形式部署，包括在计算环境中使用的独立程序、模块、组件、子程序或单元。计算机程序可以，但不是必须，对应于文件系统中的文件。程序可以存储在保存其他程序或数据的文件的一部分中，例如存储在标记语言文档中的一个或多个脚本，存储在专用于所讨论的程序的单个文件中，或者存储在存储一个或多个模块、子程序或部分代码的多个协同文件中。计算机程序可以被部署用于在一台计算机或多台计算机上执行，这些计算机例如位于一个站点或分布在通过通信网络互连的多个站点上。虽然各图中所示的部分程序可以被示为通过各种对象、方法或过程实现各种特征和功能的单独模块，但是这些程序可以替代地包括多个子模块、第三方服务、组件和库。相反，各种组件的特征和功能可以适当地组合成单个组件。用于进行计算确定的阈值可以是静态的、动态的，或者是静态和动态确定的。

[0082] 本说明书中描述的方法、过程或逻辑流程可以由执行一个或多个计算机程序的一个或多个可编程计算机来执行，以通过对输入数据进行操作并生成输出来执行功能。方法、过程或逻辑流程也可以由专用逻辑电路来执行，并且设备也可以被实现为专用逻辑电路，例如，CPU、FPGA或ASIC。

[0083] 适于执行计算机程序的计算机可以基于一个或多个通用和专用微处理器以及其他类型的CPU。计算机的元件是用于执行或执行指令的CPU和用于存储指令和数据的一个或多个存储装置。通常，CPU可以从存储器接收指令和数据(并向存储器写入数据)。计算机还可以包括或操作性地联接到一个或多个用于存储数据的大容量存储装置。在一些实现方式中，计算机可以从大容量存储装置接收数据，并将数据传输到大容量存储装置，例如包括磁盘、磁光盘或光盘。此外，计算机可以嵌入另一个装置中，例如，移动电话、个人数字助理(PDA)、移动音频或视频播放器、游戏控制台、全球定位系统(GPS)接收器或便携式存储装置

(例如,通用串行总线(USB)闪存驱动器)。

[0084] 适合于存储计算机程序指令和数据的计算机可读介质(暂时的或非暂时的,视情况而定)可以包括所有形式的永久/非永久和易失性/非易失性存储器、介质和存储装置。计算机可读介质可包括例如半导体存储器装置,诸如随机存取存储器(RAM)、只读存储器(ROM)、相变存储器(PRAM)、静态随机存取存储器(SRAM)、动态随机存取存储器(DRAM)、可擦除可编程只读存储器(EPROM)、电可擦除可编程只读存储器(EEPROM)和闪存存储器装置。计算机可读介质还可以包括例如磁性装置,例如磁带、盒式磁带、盒式磁带和内部/可移动磁盘。计算机可读介质还可包括磁光盘和光学存储器装置和技术,包括例如数字视频光盘(DVD)、CD ROM、DVD+/-R、DVD-RAM、DVD-ROM、HD-DVD和BLURAY。内存可以存储各种对象或数据,包括缓存、类、框架、应用、模块、备份数据、作业、网页、网页模板、数据结构、数据库表、存储库和动态信息。存储在内存中的对象和数据的类型可以包括参数、变量、算法、指令、规则、约束和引用。此外,内存可以包括日志、策略、安全或访问数据以及报告文件。处理器和存储器可以由专用逻辑电路补充或结合在其中。

[0085] 本公开文本中描述的主题的实现方式可以在具有显示装置的计算机上实现,该显示装置用于提供与用户的交互,包括向用户显示信息(以及从用户接收输入)。显示装置的类型可以包括,例如,阴极射线管(CRT)、液晶显示器(LCD)、发光二极管(LED)和等离子显示器。显示装置可以包括键盘和定点装置,包括例如鼠标、轨迹球或轨迹板。还可以通过使用触摸屏向计算机提供用户输入,例如具有压力灵敏度的平板计算机表面或使用电容或电传感的多点触摸屏。其他种类的装置可以用于提供与用户的交互,包括接收用户反馈,包括例如包括视觉反馈、听觉反馈或触觉反馈的感觉反馈。可以以声音、语音或触觉输入的形式接收来自用户的输入。另外,计算机可以通过向用户使用的装置发送文档和从用户使用的装置接收文档来与用户交互。例如,计算机可以响应于从网络浏览器接收的请求,向用户的客户端装置上的网络浏览器发送网页。

[0086] 术语“图形用户接口”或“GUI”可以用单数或复数来描述一个或多个图形用户接口以及特定图形用户接口的每个显示。因此,GUI可以代表任何图形用户接口,包括但不限于网络浏览器、触摸屏或命令行接口(CLI),其处理信息并向用户有效地呈现信息结果。一般来说,GUI可以包括多个用户接口(UI)元素,其中的一些或全部元素与网络浏览器相关联,例如交互字段、下拉列表和按钮。这些和其他UI元素可以与网络浏览器的功能相关或代表网络浏览器的功能。

[0087] 本说明书中描述的主题的实现方式可以在包括后端组件(例如,作为数据服务器)或包括中间件组件(例如,应用服务器)的计算系统中实现。此外,计算系统可以包括前端组件,例如,具有图形用户接口或网络浏览器之一或两者的客户端计算机,用户可以通过该图形用户接口或网络浏览器与计算机交互。系统的组件可以通过通信网络中任何形式或介质的有线或无线数字数据通信(或数据通信的组合)来互连。通信网络的例子包括局域网(LAN)、无线接入网络(RAN)、城域网(MAN)、广域网(WAN)、全球微波互联接入(WIMAX)、无线局域网(WLAN)(例如,使用802.11a/b/g/n或802.20或协议的组合)、互联网的全部或一部分或在一个或多个位置处的任何其他通信系统或系统(或通信网络的组合)。网络可以与例如互联网协议(IP)数据包、帧中继帧、异步传输模式(ATM)单元、语音、视频、数据或网络地址之间的通信类型的组合通信。

[0088] 计算系统可以包括客户端和服务器。客户端和服务器通常可以彼此远离，并且通常可以通过通信网络进行交互。客户端和服务器的关系可以借助于运行在各自计算机上并具有客户端-服务器关系的计算机程序而产生。

[0089] 集群文件系统可以是任何类型的文件系统，可从多个服务器访问以进行读取和更新。锁定或一致性跟踪可能是不必要的，因为交换文件系统的锁定可以在应用层处完成。此外，Unicode数据文件可以不同于非Unicode数据文件。

[0090] 虽然本说明书包含许多具体的实施细节，但是这些不应被解释为对所要求保护的范围的限制，而是对特定实现方式特有的特征的描述。本说明书中在分别的实现方式的上下文中描述的某些特征也可以在单个实现方式中组合实现。相反，在单个实现方式的上下文中描述的各种特征也可以在多个实现方式中分别实现，或者以任何合适的子组合实现。此外，尽管先前描述的特征可以被描述为在某些组合中起作用，并且甚至最初如此要求保护，但是在一些情况下，要求保护的组合中的一个或多个特征可以从该组合中删除，并且要求保护的组合可以指向子组合或子组合的变体。

[0091] 在前面的描述中，已经参考许多具体细节描述了本发明的实施方案，这些细节可能因实现方式而异。因此，说明书和附图被认为是说明性的，而不是限制性的。本发明范围的唯一和排他的指示、以及申请人希望的本发明的范围是本申请中以特定形式给出的权利要求(包括任何后续的修正)的字面和等同范围。本文中对包含在权利要求中的术语的任何明确定义将决定权利要求中使用的术语的含义。另外，当我们在前面的描述或后面的权利要求中使用术语“进一步包括”或“进一步包含”时，这个短语后面的可以是附加步骤或实体，或者前面叙述的步骤或实体的子步骤/子实体。

[0092] 已经描述了主题的特定实现方式。对本领域技术人员来说显而易见的是，所描述的实现方式的其他实现方式、变更和置换都在以下权利要求的范围内。虽然在附图或权利要求中以特定顺序描述了操作，但是这不应该理解为要求必须以所示的特定顺序或序列执行这些操作，或者要求执行所有示出的操作(一些操作可以被认为是可选的)，以获得期望的结果。在某些情况下，多任务或并行处理(或多任务和并行处理的组合)可能是有利的，并且被认为是适当的。

[0093] 此外，前面描述的实现方式中的各种系统模块和组件的分离或集成不应该被理解为在所有实现方式中都需要这种分离或集成，并且应该理解，所描述的程序组件和系统通常可以集成在单个软件产品中或者封装到多个软件产品中。

[0094] 因此，先前描述的示例实现方式不限定或限制本公开文本。在不脱离本公开文本的精神和范围的情况下，其他改变、替换和变更也是可能的。

[0095] 此外，任何要求保护的实现方式被认为至少适用于：计算机实现的方法；存储计算机可读指令以执行计算机实现的方法的非暂时性计算机可读介质；和计算机系统，该计算机系统包括与硬件处理器可操作地互连的计算机存储器，该硬件处理器被配置成执行计算机实现的方法或存储在非暂时性计算机可读介质上的指令。

[0096] 已经描述了这些系统和方法的多个实施方案。然而，应当理解，在不脱离本公开文本的精神和范围的情况下，可以进行各种修改。

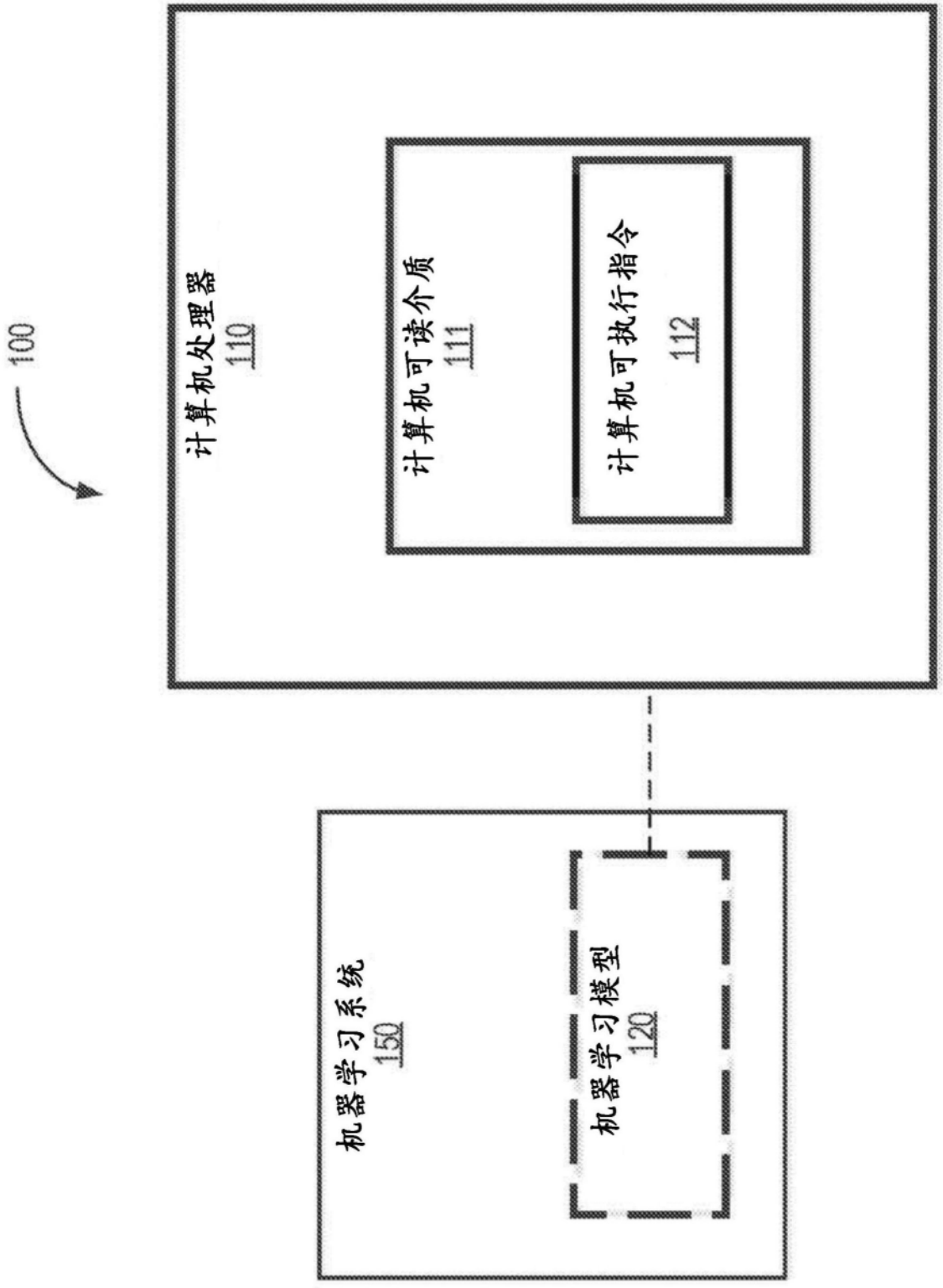

100
计算机处理器
110
计算机可读介质
111
计算机可执行指令
112
机器学习系统
150
机器学习模型
120

图1

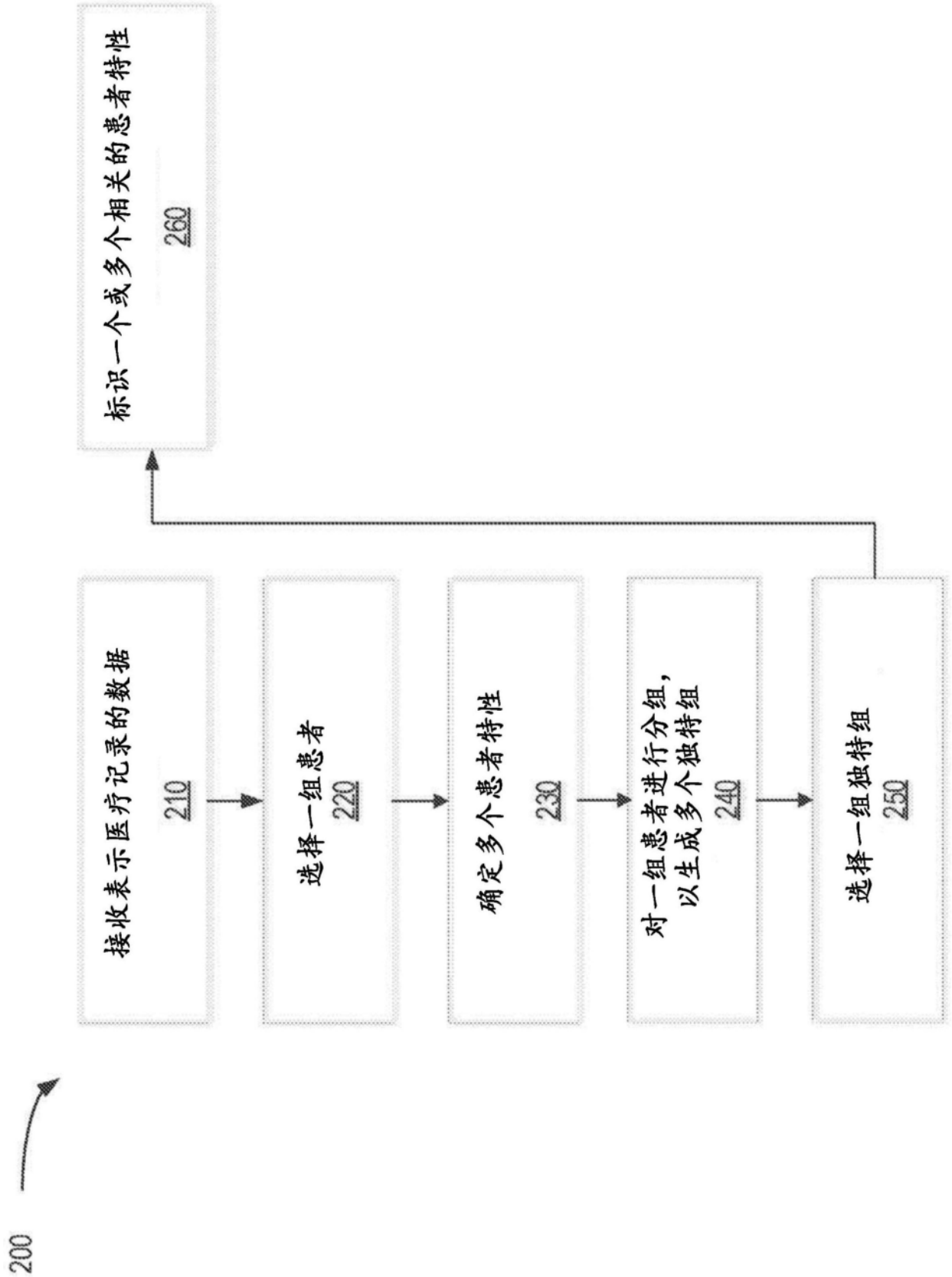

200
接收表示医疗记录的数据
210
选择一组患者
220
确定多个患者特性
230
对一组患者进行分组，
以生成多个独特组
240
选择一组独特组
250
标识一个或多个相关的患者特性
260

图2

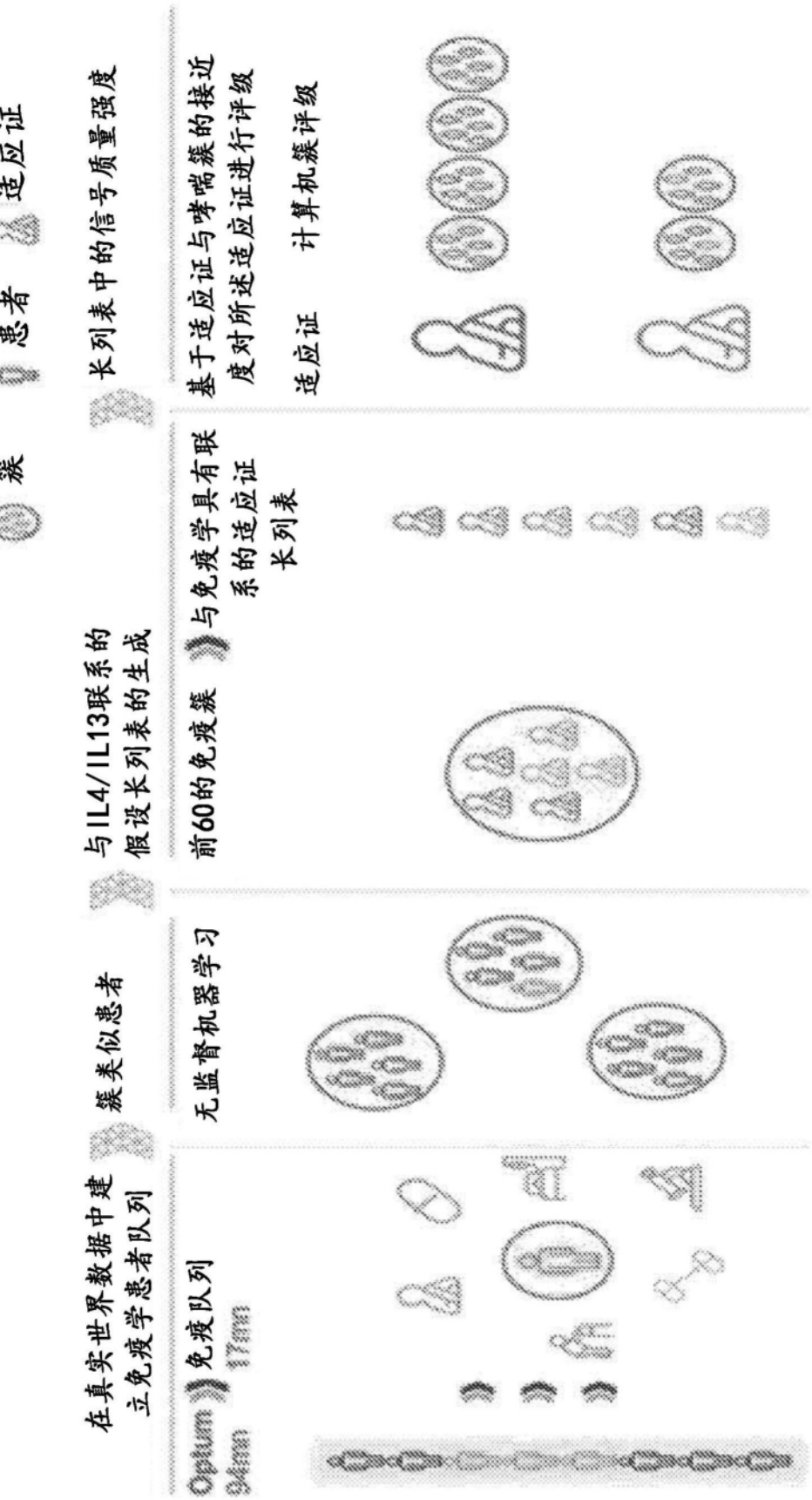

簇
患者
适应证
在真实世界数据中建立免疫学患者队列
簇类似患者
与IL4/IL13联系的假设长列表的生成
与免疫学具有联系的适应证长列表
基于适应证与哮喘簇的接近度对所述适应证进行评级
Optum
Cerner
免疫队列
无监督机器学习
前60的免疫簇
适应证
计算机簇评级

图3

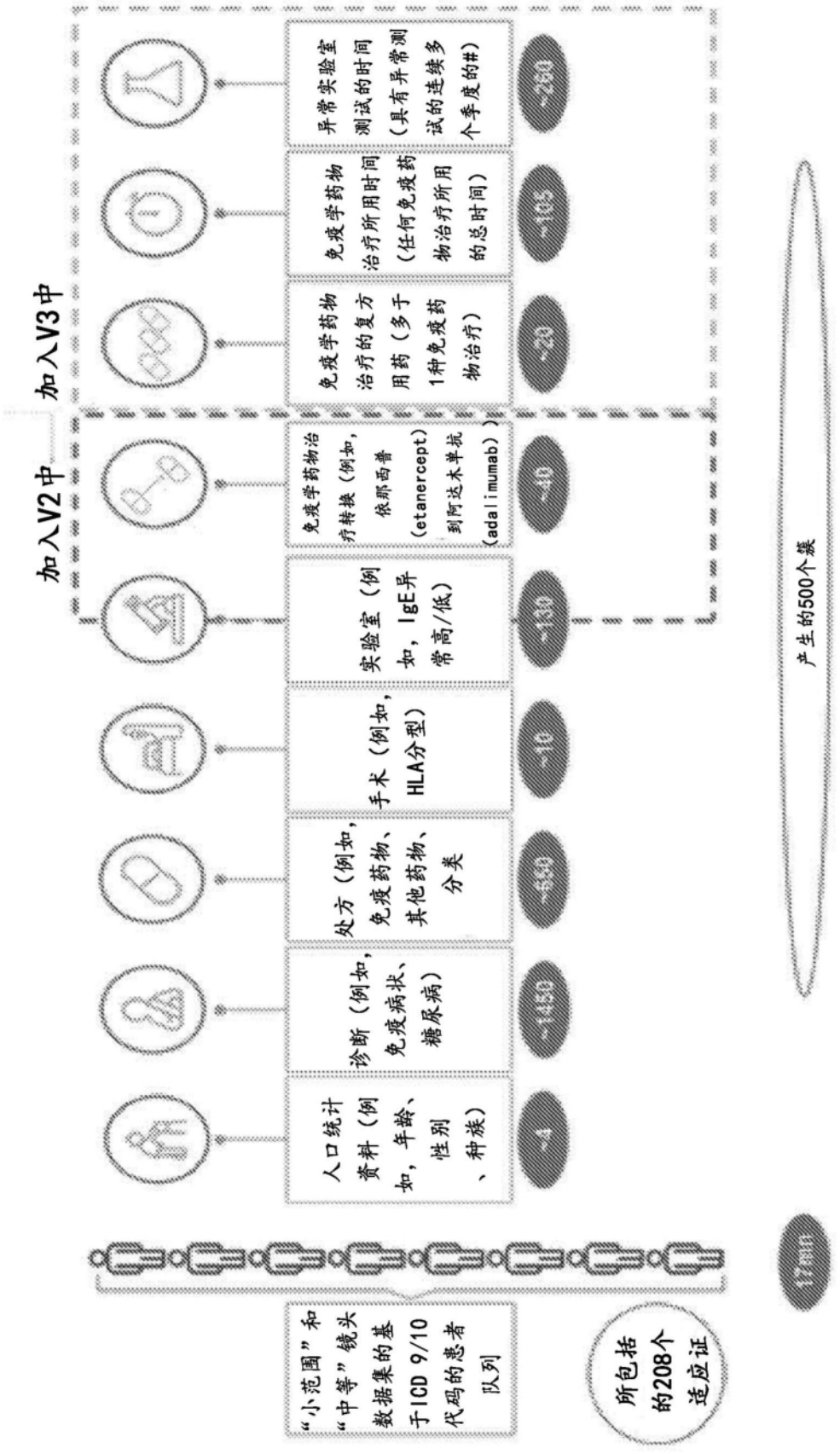

加入V2中
加入V3中
人口统计资料（例如，年龄、性别、种族）
~4
诊断（例如，免疫病状、糖尿病）
~1450
处方（例如，免疫药物、其他药物、分类
~550
手术（例如，HLA分型）
~10
实验室（例如，IgE异常高/低）
~130
免疫学药物治疗转换（例如，依那西普（etanercept）到阿达木单抗（adalimumab））
~40
免疫学药物治疗的复方用药（多于1种免疫药物治疗）
~20
免疫学药物治疗所用时间（任何免疫药物治疗所用的总时间）
~105
异常实验室测试的时间（具有异常测试的连续多个季度的#）
~260
“小范围”和“中等”镜头数据集的基于ICD 9/10代码的患者队列
所包括的208个适应证
17min
产生的500个簇

图4

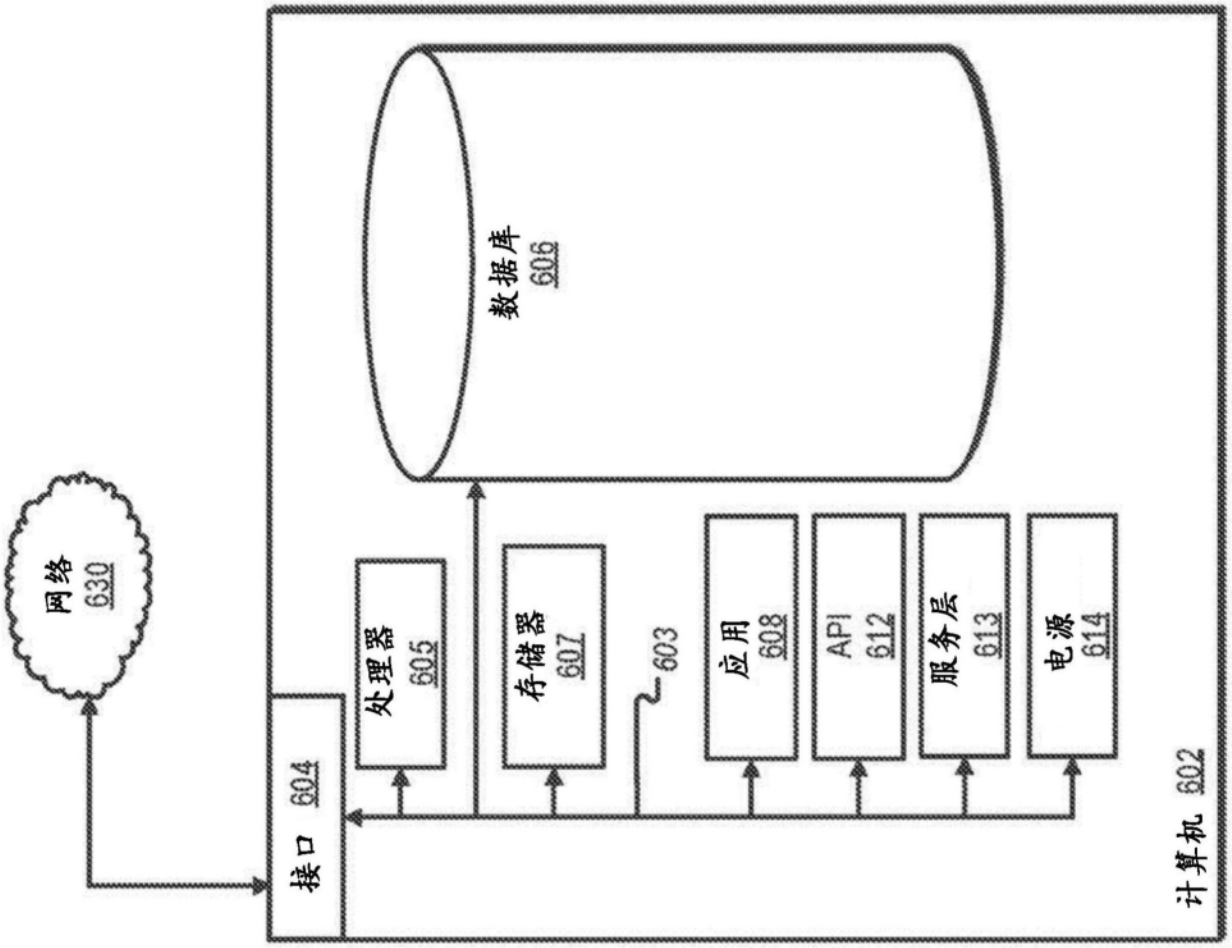

网络
630
接口 604
处理器
605
存储器
607
603
数据库
606
应用
608
API
612
服务层
613
电源
614
计算机 602

图5