

-1-

一种基因关联程度确定方法及相关设备

本申请要求于 2020 年 12 月 23 日提交中国专利局、申请号为 202011535972.2、发明名称为“一种基因关联程度确定方法及装置”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

5

技术领域

本申请属于基因关联技术领域，更具体的说，尤其涉及一种基因关联程度确定方法及相关设备。

背景技术

10 目前，随着医疗研究的不断深入，确定出与遗传性疾病有关的致病基因，将为相关人员对该遗传性疾病的研究以及后续处理提供重要的辅助作用。

遗传性疾病一般与多个致病基因有关。患有遗传性疾病的病人的病例描述文本中可能存在由疾病描述词条描述的多种临床表型信息。不同的临床表型信息关联的致病基因可能有一个或多个。不同的临床表型信息关联的致病基因可能相同，也可能不同。在当前确定的遗传性疾病有 5181 种，基因有 15428 种的情况下，如何快速确定病人的病例描述文本中的疾病描述词条与各基因的关联程度，成为相关人员急需解决的技术问题。

发明内容

20 鉴于上述问题，本申请提供一种克服上述问题或者至少部分地解决上述问题的一种基因关联程度确定方法及相关设备，相关技术方案如下：

一种基因关联程度确定方法，包括：

确定病例描述文本中的疾病描述词条；

25 对多个预设关联数据库：分别根据所述疾病描述词条与多个基因中各基因对应的基因标识，在该预设关联数据库中确定目标关联记录的记录数据，将所述记录数据输入至预设词条-基因关联矩阵中，确定所述疾病描述词条分别与所述多个基因中各基因在该预设关联数据库中的关联分数，其中，各预设关联数据库中均保存有所述疾病描述词条与所述多个基因中至少一个基因对应的基因标识的关联记录；

根据所述疾病描述词条与所述多个基因对应的基因标识分别在所述多个

-2-

预设关联数据库中的关联分数, 确定所述疾病描述词条分别与所述多个基因的关联程度。

可选的, 所述分别根据所述疾病描述词条与多个基因中各基因对应的基因标识, 在该预设关联数据库中确定目标关联记录的记录数据, 包括:

- 5 对多个基因中的任一基因: 分别确定所述疾病描述词条与该基因对应的基因标识在该预设关联数据库中的目标关联记录的记录数据, 所述目标关联记录的记录数据包括: 同时包括该基因的基因标识和所述疾病描述词条的关联记录的第一数量、包括所述疾病描述词条的关联记录的第二数量、包括该基因的基因标识的关联记录的第三数量以及该预设关联数据库中的关联记录的总数量。

- 10 可选的, 所述预设词条-基因关联矩阵为:

$$M = \frac{T_1}{T_2} \times \log \left(\frac{T_s}{T_3 + 1} \right)$$

其中, M 为在该预设信息-基因数据库中所述疾病描述词条与该基因对应的基因标识的关联分数; T_1 为所述第一数量; T_2 为所述第二数量; T_3 为所述第三数量; T_s 为所述总数量。

- 15 可选的, 所述根据所述疾病描述词条与所述多个基因对应的基因标识分别在所述多个预设关联数据库中的关联分数, 确定所述疾病描述词条分别与所述多个基因的关联程度, 包括:

对任一基因: 将所述疾病描述词条与该基因对应的基因标识在所述多个预设信息-基因关联数据库中分别对应的关联分数进行加权处理, 获得所述疾病

- 20 描述词条与该基因的关联值;

根据所述关联值, 确定所述疾病描述词条与该基因的关联程度。

可选的, 所述根据所述关联值, 确定所述疾病描述词条与该基因的关联程度, 包括:

根据

25
$$L = \frac{1}{(1 + e^{-z})}$$

确定所述疾病描述词条与该基因的关联程度, 其中, L 为所述疾病描述词

-3-

条与该基因的关联程度； e 为自然常数； z 为所述关联值。

可选的，所述预设词条-基因关联矩阵的训练过程包括：

将作为训练信息的多个疾病描述词条与所述多个基因对应的基因标识在所述多个预设信息-基因关联数据库中的目标关联记录的记录数据输入至当前的词条-基因关联训练矩阵中，确定所述训练信息与所述多个基因在各预设信息-基因数据库下分别对应的预测关联分数；

根据确定的各所述预测关联分数，分别确定所述多个疾病描述词条与所述多个基因的预测关联程度；

确定所述多个疾病描述词条与所述多个基因的真实关联程度与所述预测关联程度的误差；

判断所述误差是否处于预设误差阈值以下，如果是，则确定所述当前的词条-基因关联训练矩阵为所述词条-基因关联矩阵，如果不是，则通过梯度下降算法调整所述当前的词条-基因关联训练矩阵中的矩阵参数，将调整后的词条-基因关联训练矩阵作为当前的词条-基因关联训练矩阵，返回执行所述将作为训练信息的多个疾病描述词条与所述多个基因对应的基因标识在所述多个预设信息-基因关联数据库中的目标关联记录的记录数据输入至当前的词条-基因关联训练矩阵中，确定所述训练信息与所述多个基因在各预设信息-基因数据库下分别对应的预测关联分数的步骤。

一种基因关联程度确定装置，包括：疾病描述词条确定单元、关联分数确定单元以及关联程度确定单元，

所述疾病描述词条确定单元，用于确定病例描述文本中的疾病描述词条；

所述关联分数确定单元，用于对多个预设关联数据库：分别根据所述疾病描述词条与多个基因中各基因对应的基因标识，在该预设关联数据库中确定目标关联记录的记录数据，将所述记录数据输入至预设词条-基因关联矩阵中，确定所述疾病描述词条分别与所述多个基因中各基因在该预设关联数据库中的关联分数，其中，各预设关联数据库中均保存有所述疾病描述词条与所述多个基因中至少一个基因对应的基因标识的关联记录；

所述关联程度确定单元，用于根据所述疾病描述词条与所述多个基因对应

的基因标识分别在所述多个预设关联数据库中的关联分数,确定所述疾病描述
词条分别与所述多个基因的关联程度。

可选的,所述关联分数确定单元,具体用于对多个基因中的任一基因:分
别确定所述疾病描述词条与该基因对应的基因标识在该预设关联数据库中的
5 目标关联记录的记录数据,所述目标关联记录的记录数据包括:同时包括该基
因的基因标识和所述疾病描述词条的关联记录的第一数量、包括所述疾病描述
词条的关联记录的第二数量、包括该基因的基因标识的关联记录的第三数量以
及该预设关联数据库中的关联记录的总数量。

可选的,所述预设词条-基因关联矩阵为:

10
$$M = \frac{T_1}{T_2} \times \log \left(\frac{T_s}{T_3 + 1} \right)$$

其中, M 为在该预设信息-基因数据库中所述疾病描述词条与该基因对应
的基因标识的关联分数; T_1 为所述第一数量; T_2 为所述第二数量; T_3 为所述第
三数量; T_s 为所述总数量。

可选的,所述关联程度确定单元包括关联值获得子单元和关联程度确定子
15 单元,

所述关联值获得子单元,用于对任一基因:将所述疾病描述词条与该基因
对应的基因标识在所述多个预设信息-基因关联数据库中分别对应的关联分数
进行加权处理,获得所述疾病描述词条与该基因的关联值;

所述关联程度确定子单元,用于根据所述关联值,确定所述疾病描述词条
20 与该基因的关联程度。

可选的,所述关联程度确定子单元具体用于根据

$$L = \frac{1}{(1 + e^{-z})}$$

确定所述疾病描述词条与该基因的关联程度,其中, L 为所述疾病描述词
条与该基因的关联程度; e 为自然常数; z 为所述关联值。

25 可选的,所述装置还包括:预测关联分数确定单元、预测关联程度确定单
元、误差确定单元、误差比较单元、词条-基因关联矩阵确定单元以及矩阵参

数调整单元，

所述预测关联分数确定单元，用于将作为训练信息的多个疾病描述词条与
所述多个基因对应的基因标识在所述多个预设信息-基因关联数据库中的目标
关联记录的记录数据输入至当前的词条-基因关联训练矩阵中，确定所述训练
5 信息与所述多个基因在各预设信息-基因数据库下分别对应的预测关联分数；

所述预测关联程度确定单元，用于根据确定的各所述预测关联分数，分别
确定所述多个疾病描述词条与所述多个基因的预测关联程度；

所述误差确定单元，用于确定所述多个疾病描述词条与所述多个基因的真
实关联程度与所述预测关联程度的误差；

- 10 所述误差比较单元，用于判断所述误差是否处于预设误差阈值以下，如果
是，则触发所述词条-基因关联矩阵确定单元，如果不是，则触发所述矩阵参
数调整单元；

所述词条-基因关联矩阵确定单元，用于确定所述当前的词条-基因关联训
练矩阵为所述词条-基因关联矩阵；

- 15 所述矩阵参数调整单元，用于通过梯度下降算法调整所述当前的词条-基
因关联训练矩阵中的矩阵参数，将调整后的词条-基因关联训练矩阵作为当前
的词条-基因关联训练矩阵，触发所述预测关联分数确定单元。

一种存储介质，所述存储介质上存储有程序，该程序被处理器执行时实现
上述任一项所述的基因关联程度确定方法。

- 20 一种处理器，所述处理器用于运行程序，其中，所述程序运行时执行上述
任一项所述的基因关联程度确定方法。

一种电子设备，所述电子设备包括至少一个处理器、以及与处理器连接的
至少一个存储器、总线；其中，处理器、存储器通过总线完成相互间的通信；
处理器用于调用存储器中的程序指令，以执行上述任一项所述的基因关联程度
25 确定方法。

一种计算机程序产品，当在电子设备上执行时，适于执行初始化上述任一
项所述的基因关联程度确定方法步骤的程序。

借由上述技术方案，本申请提供一种基因关联程度确定方法及相关设

-6-

备,可以确定病例描述文本中的疾病描述词条;对多个预设关联数据库:分别根据疾病描述词条与多个基因中各基因对应的基因标识,在该预设关联数据库中确定目标关联记录的记录数据,将记录数据输入至预设词条-基因关联矩阵中,确定疾病描述词条分别与多个基因中各基因在该预设关联数据库中的关联分数,其中,各预设关联数据库中均保存有疾病描述词条与多个基因中至少一个基因对应的基因标识的关联记录;根据疾病描述词条与多个基因对应的基因标识分别在多个预设关联数据库中的关联分数,确定疾病描述词条分别与多个基因的关联程度。本发明通过在各预设关联数据库中确定疾病描述词条与多个基因的目标关联记录的记录数据,将该记录数据输入至预设词条-基因关联矩阵中,从而确定出该疾病描述词条与多个基因在各预设关联数据库中的关联分数,进而快速地获得该疾病描述词条与多个基因的关联程度。

上述说明仅是本申请技术方案的概述,为了能够更清楚了解本申请的技术手段,而可依照说明书的内容予以实施,并且为了让本申请的上述和其它目的、特征和优点能够更明显易懂,以下特举本申请的具体实施方式。

15 附图说明

为了更清楚地说明本申请实施例或现有技术中的技术方案,下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍,显而易见地,下面描述中的附图是本申请的一些实施例,对于本领域普通技术人员来讲,在不付出创造性劳动的前提下,还可以根据这些附图获得其他的附图。

20 图 1 示出了本发明实施例提供的一种基因关联程度确定方法的流程示意图;

图 2 示出了本发明实施例提供的另一种基因关联程度确定方法的流程示意图;

25 图 3 示出了本发明实施例提供的预设词条-基因关联矩阵的训练过程示意图;

图 4 示出了本发明实施例提供的另一种基因关联程度确定方法的流程示意图;

图 5 示出了本发明实施例提供的一种基因关联程度确定装置的结构示意

图；

图 6 示出了本发明实施例提供的另一种基因关联程度确定装置的结构示意图。

具体实施方式

5 下面将参照附图更详细地描述本公开的示例性实施例。虽然附图中显示了本公开的示例性实施例，然而应当理解，可以以各种形式实现本公开而不应被这里阐述的实施例所限制。相反，提供这些实施例是为了能够更透彻地理解本公开，并且能够将本公开的范围完整的传达给本领域的技术人员。

如图 1 所示，本发明实施例提供的一种基因关联程度确定方法，可以包括：

10 S100、确定病例描述文本中的疾病描述词条。

其中，病例描述文本可以是医务人员对患者的疾病进行检查、诊断等医疗活动过程的记录。可选的，病例描述文本可以由多个词汇组成的句子或段落。疾病描述词条可以是对患者的疾病的临床表型信息进行描述的词条。

可选的，该疾病描述词条可以为 HPO 词条或遗传性疾病词条。其中，HPO
15 (Human Phenotype Ontology, 人类表型标准用语) 是一套描述人类疾病所致的异常表型的标准词汇集，该标准词汇集包括多个 HPO 词条。可以理解的是，HPO 词条可以是中文词汇，也可以是其他语言的词汇，例如：英语词汇等。遗传性疾病词条可以是以基因为主要致病原因的疾病的专业名称、和/或简称、和/或其他自定义名称。需要注意的是，该其他自定义名称可以是业内对该疾病的口语化表达词汇。例如：专业名称为“支气管哮喘”的疾病，简称可以是
20 “哮喘病”，在口语化中也经常使用“喘息症”进行表达。可以理解的是，由于遗传性疾病的在研究领域的不断深入以及在各应用领域的实际需求可能不同，业内人员可以根据研究需要或实际应用需要，自行创建遗传性疾病词条。

可选的，本发明实施例可以对病例描述文本进行分词处理，再通过自然语
25 言处理 (Natural Language Processing, NLP) 中的命名实体识别 (Named Entity Recognition, NER)，从非结构化的病例描述文本中提取出结构化的疾病描述词条。本发明实施例通过使用命名实体识别技术，可以从病例描述文本中准确提取出疾病描述词条。

可以理解的是,本发明实施例也可以通过预设疾病描述关键词在病例描述文本中检索出疾病描述词条。

可以理解的是,病例描述文本中可以包括一个或多个疾病描述词条,本发明实施例可以对从该病例描述文本确定出的任一疾病描述词条执行本发明实
5 施例提供的基因关联程度确定方法。

S200、对多个预设关联数据库:分别根据该疾病描述词条与多个基因中各基因对应的基因标识,在该预设关联数据库中确定目标关联记录的记录数据,将该记录数据输入至预设词条-基因关联矩阵中,确定该疾病描述词条分别与
10 该多个基因中各基因在该预设关联数据库中的关联分数,其中,各预设关联数据库
中均保存有该疾病描述词条与该多个基因中至少一个基因对应的基因标识的关联记录。

预设关联数据库可以是基于业内公共疾病数据库,构建的可映射词条与基因标识关联的数据库。可选的,业内公共疾病数据库可以包括: OMIM (疾病相关基因) 数据库、ClinVar (遗传变异) 数据库、HGMD (人类基因突变)
15 数据库、HPO (人类表型标准用语) 数据库以及 Orphanet (稀有疾病相关) 数据库。本发明实施例可以对不同的业内公共疾病数据库分别构建关联数据库。可选的,预设关联数据库中
可以保存有映射疾病描述词条与基因标识之间关联关系的关联记录。

可选的,本发明实施例可以根据疾病描述词条与多个基因对应的基因标识,
20 在预设关联数据库中确定出目标关联记录的记录数据。可选的,目标关联记录的记录数据
可以是在该预设关联数据库中同时包括该疾病描述词条与该多个基因中的任一基因的
关联记录的数量,也可以是在该预设关联数据库中包括该多个基因中的任一基因的基因
标识的关联记录的数量。

可选的,步骤 S200 可以包括:对多个基因中的任一基因:分别确定该疾
25 病描述词条与该基因对应的基因标识在该预设关联数据库中的目标关联记录的记录数据,
该目标关联记录的记录数据包括:同时包括该基因的基因标识和该疾病描述词条的关联
记录的第一数量、包括该疾病描述词条的关联记录的第二数量、包括该基因的基因标识
的关联记录的第三数量以及该预设关联数据库

中的关联记录的总数量。

可以理解的是，目标关联记录可以是该预设关联数据库中的关联记录，可以是在该预设关联数据库中同时包括该基因的基因标识和该疾病描述词条的关联记录，可以是在该预设关联数据库中包括该疾病描述词条的关联记录，可以是在该预设关联数据库中包括该基因的基因标识的关联记录。

5 以是包括该基因的基因标识的关联记录。

可选的，该预设词条-基因关联矩阵可以为：

$$M=\frac{T_1}{T_2}\times\log\left(\frac{T_s}{T_3+1}\right)$$

其中， M 为在该预设信息-基因数据库中该疾病描述词条与该基因对应的基因标识的关联分数； T_1 为该第一数量； T_2 为该第二数量； T_3 为该第三数量；

10 T_s 为该总数量。

本发明实施例通过将确定的目标关联记录的记录数据输入预设词条-基因关联矩阵中进行计算，可以获得预设词条-基因关联矩阵输出的关联分数。需要注意的是，确定的该疾病描述词条与该多个基因中的任一基因的关联分数的数量与该多个预设关联数据库的数量相同。例如：假设该多个预设关联数据库有五个，对于该疾病描述词条和任一基因：本发明实施例可以通过预设词条-基因关联矩阵，分别在这五个预设关联数据库中确定出该疾病描述词条和该基因的关联分数，即确定出该疾病描述词条和该基因五个关联分数。可以理解的是，该五个关联分数可以相同，也可以不同。

15 本发明实施例可以通过预设词条-基因关联矩阵，分别在这五个预设关联数据库中确定出该疾病描述词条和该基因的关联分数，即确定出该疾病描述词条和该基因五个关联分数。可以理解的是，该五个关联分数可以相同，也可以不同。

S300、根据该疾病描述词条与该多个基因对应的基因标识分别在该多个预设关联数据库中的关联分数，确定该疾病描述词条分别与该多个基因的关联程度。

可选的，本发明实施例可以对任一基因：将该疾病描述词条与该基因对应的基因标识在该多个预设信息-基因关联数据库中分别对应的关联分数进行相加求和，获得该疾病描述词条与该基因的关联值。

25 可选的，基于图1所示的方法，如图2所示，本发明实施例提供的另一种基因关联程度确定方法，步骤S300可以包括：

S310、对任一基因：将该疾病描述词条与该基因对应的基因标识在该多个

预设信息-基因关联数据库中分别对应的关联分数进行加权处理，获得该疾病描述词条与该基因的关联值。

本发明实施例可以预先对各个预设信息-基因关联数据库设置权重，使得经该加权处理后获得的该疾病描述词条与该基因的关联值更加可靠。为了便于理解，此处通过举例进行说明：预先设置预设信息-基因关联数据库 A 的权重为 w_1 、预设信息-基因关联数据库 B 的权重为 w_2 、预设信息-基因关联数据库 C 的权重为 w_3 、预设信息-基因关联数据库 D 的权重为 w_4 以及预设信息-基因关联数据库 E 的权重为 w_5 ，该疾病描述词条与该基因对应的基因标识在预设信息-基因关联数据库 A 中对应的关联分数为 x_1 、在预设信息-基因关联数据库 B 中对应的关联分数为 x_2 、在预设信息-基因关联数据库 C 中对应的关联分数为 x_3 、在预设信息-基因关联数据库 D 中对应的关联分数为 x_4 以及在预设信息-基因关联数据库 E 中对应的关联分数为 x_5 ，则通过加权处理，获得该疾病描述词条与该基因的关联值为：

$$z = x_1 w_1 + x_2 w_2 + x_3 w_3 + x_4 w_4 + x_5 w_5。$$

S320、根据该关联值，确定该疾病描述词条与该基因的关联程度。

可选的，步骤 S320 可以包括：

根据

$$L = \frac{1}{(1 + e^{-z})}$$

确定该疾病描述词条与该基因的关联程度，其中， L 为该疾病描述词条与该基因的关联程度； e 为自然常数； z 为该关联值。

本发明提供的一种基因关联程度确定方法，可以确定病例描述文本中的疾病描述词条；对多个预设关联数据库：分别根据疾病描述词条与多个基因中各基因对应的基因标识，在该预设关联数据库中确定目标关联记录的记录数据，将记录数据输入至预设词条-基因关联矩阵中，确定疾病描述词条分别与多个基因中各基因在该预设关联数据库中的关联分数，其中，各预设关联数据库中均保存有疾病描述词条与多个基因中至少一个基因对应的基因标识的关联记录；根据疾病描述词条与多个基因对应的基因标识分别在多个预设关联数据库

中的关联分数，确定疾病描述词条分别与多个基因的关联程度。本发明通过在各预设关联数据库中确定疾病描述词条与多个基因的目标关联记录的记录数据，将该记录数据输入至预设词条-基因关联矩阵中，从而确定出该疾病描述词条与多个基因在各预设关联数据库中的关联分数，进而快速地获得该疾病描述词条与多个基因的关联程度。

可选的，如图 3 所示，本发明实施例提供的预设词条-基因关联矩阵的训练过程可以包括：

S10、将作为训练信息的多个疾病描述词条与该多个基因对应的基因标识在该多个预设信息-基因关联数据库中的目标关联记录的记录数据输入至当前的词条-基因关联训练矩阵中，确定该训练信息与该多个基因在各预设信息-基因数据库下分别对应的预测关联分数。

本发明实施例可以在训练刚开始时，初始化词条-基因关联训练矩阵。步骤 S10 与步骤 S200 相似，本发明实施例可以依次将多个疾病描述词条与该多个基因对应的基因标识在该多个预设信息-基因关联数据库中的目标关联记录的记录数据输入至当前的词条-基因关联训练矩阵中，获得当前的词条-基因关联训练矩阵输出的预测关联分数。同理，确定的该疾病描述词条与该多个基因中的任一基因的预测关联分数的数量与该多个预设关联数据库的数量相同。

S20、根据确定的各该预测关联分数，分别确定该多个疾病描述词条与该多个基因的预测关联程度。

可选的，本发明实施例可以对任一基因：将该疾病描述词条与该基因对应的基因标识在该多个预设信息-基因关联数据库中分别对应的预测关联分数进行相加求和，获得该疾病描述词条与该基因的预测关联值。

可选的，本发明实施例可以对任一基因：将该疾病描述词条与该基因对应的基因标识在该多个预设信息-基因关联数据库中分别对应的预测关联分数进行加权处理，获得该疾病描述词条与该基因的预测关联值。

本发明实施例可以根据预测关联值，确定该疾病描述词条与该基因的预测关联程度。

S30、确定该多个疾病描述词条与该多个基因的真实关联程度与该预测关

联程度的误差。

其中,真实关联程度可以是本领域专业人员对该多个疾病描述词条与该多个基因确定的关联程度。本发明实施例可以通过真实关联程度与预测关联程度的误差,判断当前的词条-基因关联训练矩阵是否训练完成。

- 5 S40、判断该误差是否处于预设误差阈值以下,如果是,则执行步骤 S50,如果不是,则执行步骤 S60。

本发明实施例可以根据实际需要设置预设误差阈值,在此不作进一步地限制。

S50、确定该当前的词条-基因关联训练矩阵为该词条-基因关联矩阵。

- 10 S60、通过梯度下降算法调整该当前的词条-基因关联训练矩阵中的矩阵参数,将调整后的词条-基因关联训练矩阵作为当前的词条-基因关联训练矩阵,返回执行步骤 S10。

- 本发明实施例通过梯度下降算法调整矩阵参数,可以重新对调整矩阵参数后的词条-基因关联训练矩阵再一次进行验证训练,使得最终获得的词条-基因
15 关联矩阵可以获得更加准确的关联分数,进而缩小真实关联程度与该预测关联程度的误差。

- 可以理解的是,本发明实施例还可以在判断该误差不处于预设误差阈值以下时,通过梯度下降法调整各个预设信息-基因关联数据库的权重。通过调整各个预设信息-基因关联数据库的权重,可以缩小真实关联程度与该预测关联
20 程度的误差。

可选的,基于图 1 所示的方法,如图 4 所示,本发明实施例提供的另一种基因关联程度确定方法,在步骤 S300 之后,该方法还可以包括:

S400、按照该疾病描述词条分别与该多个基因的关联程度由高至低对该多个基因进行排序。

- 25 可以理解的是,基因与该疾病描述词条的关联程度越高,说明该疾病描述词条描述的疾病的临床表型与该基因的表型越相似。在确定该疾病描述词条与该多个基因的关联程度之后,对该多个基因进行排序有助于相关技术人员辨别筛选出该疾病描述词条对应的疾病的致病基因。

与上述方法实施例相对应,本发明实施例还提供一种基因关联程度确定装置,其结构如图5所示,可以包括:疾病描述词条确定单元100、关联分数确定单元200以及关联程度确定单元300。

5 该疾病描述词条确定单元100,用于确定病例描述文本中的疾病描述词条。

其中,病例描述文本可以是医务人员对患者的疾病进行检查、诊断等医疗活动过程的记录。可选的,病例描述文本可以由多个词汇组成的句子或段落。疾病描述词条可以是对患者的疾病的临床表型信息进行描述的词条。

10 可选的,该疾病描述词条可以为HPO词条或遗传性疾病词条。其中,HPO (Human Phenotype Ontology,人类表型标准用语)是一套描述人类疾病所致的异常表型的标准词汇集,该标准词汇集包括多个HPO词条。可以理解的是,HPO词条可以是中文词汇,也可以是其他语言的词汇,例如:英语词汇等。遗传性疾病词条可以是以基因为主要致病原因的疾病的专业名称、和/或简称、和/或其他自定义名称。需要注意的是,该其他自定义名称可以是业内对该疾

15 病的口语化表达词汇。

例如:专业名称为“支气管哮喘”的疾病,简称可以是“哮喘病”,在口语化中也经常使用“喘息症”进行表达。可以理解的是,由于遗传性疾病的在研究领域的不断深入以及在各应用领域的实际需求可能不同,业内人员可以根据研究需要或实际应用需要,自行创建遗传性疾病词条。

20 可选的,疾病描述词条确定单元100可以对病例描述文本进行分词处理,再通过自然语言处理(Natural Language Processing, NLP)中的命名实体识别(Named Entity Recognition, NER),从非结构化的病例描述文本中提取出结构化的疾病描述词条。本发明实施例通过使用命名实体识别技术,可以从病例描述文本中准确提取出疾病描述词条。

25 可以理解的是,疾病描述词条确定单元100也可以通过预设疾病描述关键词在病例描述文本中检索出疾病描述词条。

该关联分数确定单元200,用于对多个预设关联数据库:分别根据该疾病描述词条与多个基因中各基因对应的基因标识,在该预设关联数据库中确定目

标关联记录的记录数据，将该记录数据输入至预设词条-基因关联矩阵中，确定该疾病描述词条分别与该多个基因中各基因在该预设关联数据库中的关联分数，其中，各预设关联数据库中均保存有该疾病描述词条与该多个基因中至少一个基因对应的基因标识的关联记录。

- 5 预设关联数据库可以是基于业内公共疾病数据库，构建的可映射词条与基因标识关联的数据库。可选的，业内公共疾病数据库可以包括：OMIM（疾病相关基因）数据库、ClinVar（遗传变异）数据库、HGMD（人类基因突变）数据库、HPO（人类表型标准用语）数据库以及 Orphanet（稀有疾病相关）数据库。可选的，预设关联数据库中保存有映射疾病描述词条与基因标识之
- 10 间关联关系的关联记录。

- 可选的，关联分数确定单元 200 可以根据疾病描述词条与多个基因对应的基因标识，在预设关联数据库中确定出目标关联记录的记录数据。可选的，目标关联记录的记录数据可以是在该预设关联数据库中同时包括该疾病描述词条与该多个基因中的任一基因的关联记录的数量，也可以是在该预设关联数据库中包括该多个基因中的任一基因的基因标识的关联记录的数量。
- 15 可选的，该关联分数确定单元 200，可以具体用于对多个基因中的任一基因：分别确定该疾病描述词条与该基因对应的基因标识在该预设关联数据库中的目标关联记录的记录数据，该目标关联记录的记录数据包括：同时包括该基因的基因标识和该疾病描述词条的关联记录的第一数量、包括该疾病描述词条
- 20 的关联记录的第二数量、包括该基因的基因标识的关联记录的第三数量以及该预设关联数据库中的关联记录的总数量。

- 可以理解的是，目标关联记录可以是该预设关联数据库中的关联记录，可以是在该预设关联数据库中同时包括该基因的基因标识和该疾病描述词条的关联记录，可以是在该预设关联数据库中包括该疾病描述词条的关联记录，可
- 25 以是包括该基因的基因标识的关联记录。

 可选的，该预设词条-基因关联矩阵可以为：

$$M = \frac{T_1}{T_2} \times \log \left(\frac{T_s}{T_3 + 1} \right)$$

其中， M 为在该预设信息-基因数据库中该疾病描述词条与该基因对应的基因标识的关联分数； T_1 为该第一数量； T_2 为该第二数量； T_3 为该第三数量； T_s 为该总数量。

关联分数确定单元 200 用于通过将确定的目标关联记录的记录数据输入预设词条-基因关联矩阵中进行计算，可以获得预设词条-基因关联矩阵输出的关联分数。需要注意的是，确定的该疾病描述词条与该多个基因中的任一基因的关联分数的数量与该多个预设关联数据库的数量相同。

该关联程度确定单元 300，用于根据该疾病描述词条与该多个基因对应的基因标识分别在该多个预设关联数据库中的关联分数，确定该疾病描述词条分别与该多个基因的关联程度。

可选的，关联程度确定单元 300 可以用于对任一基因：将该疾病描述词条与该基因对应的基因标识在该多个预设信息-基因关联数据库中分别对应的关联分数进行相加求和，获得该疾病描述词条与该基因的关联值。

可选的，该关联程度确定单元 300 包括关联值获得子单元和关联程度确定子单元。

该关联值获得子单元，用于对任一基因：将该疾病描述词条与该基因对应的基因标识在该多个预设信息-基因关联数据库中分别对应的关联分数进行加权处理，获得该疾病描述词条与该基因的关联值。

本发明实施例可以预先对各个预设信息-基因关联数据库设置权重，使得经该加权处理后获得的该疾病描述词条与该基因的关联值更加可靠。

该关联程度确定子单元，用于根据该关联值，确定该疾病描述词条与该基因的关联程度。

可选的，该关联程度确定子单元具体用于根据

$$L = \frac{1}{(1 + e^{-z})}$$

确定该疾病描述词条与该基因的关联程度，其中， L 为该疾病描述词条与该基因的关联程度； e 为自然常数； z 为该关联值。

本发明提供一种基因关联程度确定装置，可以确定病例描述文本中的疾

病描述词条；对多个预设关联数据库：分别根据疾病描述词条与多个基因中各基因对应的基因标识，在该预设关联数据库中确定目标关联记录的记录数据，将记录数据输入至预设词条-基因关联矩阵中，确定疾病描述词条分别与多个基因中各基因在该预设关联数据库中的关联分数，其中，各预设关联数据库中均保存有疾病描述词条与多个基因中至少一个基因对应的基因标识的关联记录；根据疾病描述词条与多个基因对应的基因标识分别在多个预设关联数据库中的关联分数，确定疾病描述词条分别与多个基因的关联程度。本发明通过在各预设关联数据库中确定疾病描述词条与多个基因的目标关联记录的记录数据，将该记录数据输入至预设词条-基因关联矩阵中，从而确定出该疾病描述词条与多个基因在各预设关联数据库中的关联分数，进而快速地获得该疾病描述词条与多个基因的关联程度。

可选的，基因关联程度确定装置还可以包括：预测关联分数确定单元、预测关联程度确定单元、误差确定单元、误差比较单元、词条-基因关联矩阵确定单元以及矩阵参数调整单元。

该预测关联分数确定单元，用于将作为训练信息的多个疾病描述词条与该多个基因对应的基因标识在该多个预设信息-基因关联数据库中的目标关联记录的记录数据输入至当前的词条-基因关联训练矩阵中，确定该训练信息与该多个基因在各预设信息-基因数据库下分别对应的预测关联分数。

该预测关联程度确定单元，用于根据确定的各该预测关联分数，分别确定该多个疾病描述词条与该多个基因的预测关联程度。

可选的，预测关联程度确定单元可以用于对任一基因：将该疾病描述词条与该基因对应的基因标识在该多个预设信息-基因关联数据库中分别对应的预测关联分数进行相加求和，获得该疾病描述词条与该基因的预测关联值。

可选的，预测关联程度确定单元可以用于对任一基因：将该疾病描述词条与该基因对应的基因标识在该多个预设信息-基因关联数据库中分别对应的预测关联分数进行加权处理，获得该疾病描述词条与该基因的预测关联值。

预测关联程度确定单元可以用于根据预测关联值，确定该疾病描述词条与该基因的预测关联程度。

该误差确定单元,用于确定该多个疾病描述词条与该多个基因的真实关联程度与该预测关联程度的误差。

其中,真实关联程度可以是本领域专业人员对该多个疾病描述词条与该多个基因确定的关联程度。

- 5 该误差比较单元,用于判断该误差是否处于预设误差阈值以下,如果是,则触发该词条-基因关联矩阵确定单元,如果不是,则触发该矩阵参数调整单元。

该词条-基因关联矩阵确定单元,用于确定该当前的词条-基因关联训练矩阵为该词条-基因关联矩阵。

- 10 该矩阵参数调整单元,用于通过梯度下降算法调整该当前的词条-基因关联训练矩阵中的矩阵参数,将调整后的词条-基因关联训练矩阵作为当前的词条-基因关联训练矩阵,触发该预测关联分数确定单元 200。

可选的,基于图 5 所示的装置,如图 6 所示,本发明实施例提供的另一种基因关联程度确定装置,还可以包括:基因排序单元 400。

- 15 该基因排序单元 400,可以用于在关联程度确定单元 300 确定疾病描述词条分别与多个基因的关联程度之后,按照该疾病描述词条分别与该多个基因的关联程度由高至低对该多个基因进行排序。

- 20 该基因关联程度确定装置包括处理器和存储器,上述疾病描述词条确定单元 100、关联分数确定单元 200 以及关联程度确定单元 300 等均作为程序单元存储在存储器中,由处理器执行存储在存储器中的上述程序单元来实现相应的功能。

- 25 处理器中包含内核,由内核去存储器中调取相应的程序单元。内核可以设置一个或以上,通过调整内核参数来通过各个预设关联数据库中确定疾病描述词条与多个基因的目标关联记录的记录数据,将该记录数据输入至预设词条-基因关联矩阵中,从而确定出该疾病描述词条与多个基因在各预设关联数据库中的关联分数,进而快速地获得该疾病描述词条与多个基因的关联程度。

本发明实施例提供了一种存储介质,其上存储有程序,该程序被处理器执行时实现所述基因关联程度确定方法。

本发明实施例提供了一种处理器，所述处理器用于运行程序，其中，所述程序运行时执行所述基因关联程度确定方法。

本发明实施例提供了一种电子设备，电子设备包括至少一个处理器、以及与处理器连接的至少一个存储器、总线；其中，处理器、存储器通过总线完成相互间的通信；处理器用于调用存储器中的程序指令，以执行上述的基因关联程度确定方法。本文中的电子设备可以是服务器、PC、PAD、手机等。

本申请还提供了一种计算机程序产品，当在电子设备上执行时，适于执行初始化有所述基因关联程度确定方法步骤的程序。

本申请是参照根据本申请实施例的方法、装置、电子设备（系统）、和计算机程序产品的流程图和/或方框图来描述的。应理解可由计算机程序指令实现流程图和/或方框图中的每一流程和/或方框、以及流程图和/或方框图中的流程和/或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程设备的处理器以产生一个机器，使得通过计算机或其他可编程设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的装置。

在一个典型的配置中，电子设备包括一个或多个处理器(CPU)、存储器和总线。电子设备还可以包括输入/输出接口、网络接口等。

存储器可能包括计算机可读介质中的非永久性存储器，随机存取存储器(RAM)和/或非易失性内存等形式，如只读存储器(ROM)或闪存(flash RAM)，存储器包括至少一个存储芯片。存储器是计算机可读介质的示例。

计算机可读介质包括永久性和非永久性、可移动和非可移动媒体可以由任何方法或技术来实现信息存储。信息可以是计算机可读指令、数据结构、程序的模块或其他数据。计算机的存储介质的例子包括，但不限于相变内存(PRAM)、静态随机存取存储器(SRAM)、动态随机存取存储器(DRAM)、其他类型的随机存取存储器(RAM)、只读存储器(ROM)、电可擦除可编程只读存储器(EEPROM)、快闪记忆体或其他内存技术、只读光盘只读存储器(CD-ROM)、数字多功能光盘(DVD)或其他光学存储、磁盒式磁带，磁带磁磁盘存储或其他磁性存储设备或任何其他非传输介质，可用于存储可以被计算设

备访问的信息。按照本文中的界定，计算机可读介质不包括暂存电脑可读媒体 (transitory media)，如调制的数据信号和载波。

还需要说明的是，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法、商品或者设备不仅包
5 括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、方法、商品或者设备所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括要素的过程、方法、商品或者设备中还存在另外的相同要素。

本领域技术人员应明白，本申请的实施例可提供为方法、系统或计算机程
10 序产品。因此，本申请可采用完全硬件实施例、完全软件实施例或结合软件和硬件方面的实施例的形式。而且，本申请可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质(包括但不限于磁盘存储器、CD-ROM、光学存储器等)上实施的计算机程序产品的形式。

以上仅为本申请的实施例而已，并不用于限制本申请。对于本领域技术人
15 员来说，本申请可以有各种更改和变化。凡在本申请的精神和原理之内所作的任何修改、等同替换、改进等，均应包含在本申请的权利要求范围之内。

-20-

权 利 要 求

1、一种基因关联程度确定方法，其特征在于，包括：

确定病例描述文本中的疾病描述词条；

- 对多个预设关联数据库：分别根据所述疾病描述词条与多个基因中各基因
- 5 对应的基因标识，在该预设关联数据库中确定目标关联记录的记录数据，将所述记录数据输入至预设词条-基因关联矩阵中，确定所述疾病描述词条分别与
- 所述多个基因中各基因在该预设关联数据库中的关联分数，其中，各预设关联数据库中均保存有所述疾病描述词条与所述多个基因中至少一个基因对应的
- 基因标识的关联记录；
- 10 根据所述疾病描述词条与所述多个基因对应的基因标识分别在所述多个预设关联数据库中的关联分数，确定所述疾病描述词条分别与所述多个基因的关联程度。

- 2、根据权利要求 1 所述的方法，其特征在于，所述分别根据所述疾病描述词条与多个基因中各基因对应的基因标识，在该预设关联数据库中确定目标
- 15 关联记录的记录数据，包括：

- 对多个基因中的任一基因：分别确定所述疾病描述词条与该基因对应的基因标识在该预设关联数据库中的目标关联记录的记录数据，所述目标关联记录的
- 记录数据包括：同时包括该基因的基因标识和所述疾病描述词条的关联记录的第一数量、包括所述疾病描述词条的关联记录的第二数量、包括该基因的基
- 20 因标识的关联记录的第三数量以及该预设关联数据库中的关联记录的总数量。

3、根据权利要求 2 所述的方法，其特征在于，所述预设词条-基因关联矩阵为：

$$M = \frac{T_1}{T_2} \times \log \left(\frac{T_s}{T_3 + 1} \right)$$

- 其中， M 为在该预设信息-基因数据库中所述疾病描述词条与该基因对应的
- 25 的基因标识的关联分数； T_1 为所述第一数量； T_2 为所述第二数量； T_3 为所述第三数量； T_s 为所述总数量。

4、根据权利要求 1 所述的方法，其特征在于，所述根据所述疾病描述词

条与所述多个基因对应的基因标识分别在所述多个预设关联数据库中的关联分数，确定所述疾病描述词条分别与所述多个基因的关联程度，包括：

对任一基因：将所述疾病描述词条与该基因对应的基因标识在所述多个预设信息-基因关联数据库中分别对应的关联分数进行加权处理，获得所述疾病

5 描述词条与该基因的关联值；

根据所述关联值，确定所述疾病描述词条与该基因的关联程度。

5、根据权利要求 4 所述的方法，其特征在于，所述根据所述关联值，确定所述疾病描述词条与该基因的关联程度，包括：

根据

$$10 \quad L = \frac{1}{(1 + e^{-z})}$$

确定所述疾病描述词条与该基因的关联程度，其中， L 为所述疾病描述词条与该基因的关联程度； e 为自然常数； z 为所述关联值。

6、根据权利要求 1 所述的方法，其特征在于，所述预设词条-基因关联矩阵的训练过程包括：

15 将作为训练信息的多个疾病描述词条与所述多个基因对应的基因标识在所述多个预设信息-基因关联数据库中的目标关联记录的记录数据输入至当前的词条-基因关联训练矩阵中，确定所述训练信息与所述多个基因在各预设信息-基因数据库下分别对应的预测关联分数；

20 根据确定的各所述预测关联分数，分别确定所述多个疾病描述词条与所述多个基因的预测关联程度；

确定所述多个疾病描述词条与所述多个基因的真实关联程度与所述预测关联程度的误差；

25 判断所述误差是否处于预设误差阈值以下，如果是，则确定所述当前的词条-基因关联训练矩阵为所述词条-基因关联矩阵，如果不是，则通过梯度下降算法调整所述当前的词条-基因关联训练矩阵中的矩阵参数，将调整后的词条-基因关联训练矩阵作为当前的词条-基因关联训练矩阵，返回执行所述将作为训练信息的多个疾病描述词条与所述多个基因对应的基因标识在所述多个预

设信息-基因关联数据库中的目标关联记录的记录数据输入至当前的词条-基因关联训练矩阵中，确定所述训练信息与所述多个基因在各预设信息-基因数据库下分别对应的预测关联分数的步骤。

- 7、一种基因关联程度确定装置，其特征在于，包括：疾病描述词条确定
5 单元、关联分数确定单元以及关联程度确定单元，

所述疾病描述词条确定单元，用于确定病例描述文本中的疾病描述词条；

- 所述关联分数确定单元，用于对多个预设关联数据库：分别根据所述疾病
描述词条与多个基因中各基因对应的基因标识，在该预设关联数据库中确定目
标关联记录的记录数据，将所述记录数据输入至预设词条-基因关联矩阵中，
10 确定所述疾病描述词条分别与所述多个基因中各基因在该预设关联数据库中的
关联分数，其中，各预设关联数据库中均保存有所述疾病描述词条与所述多
个基因中至少一个基因对应的基因标识的关联记录；

- 所述关联程度确定单元，用于根据所述疾病描述词条与所述多个基因对应的
基因标识分别在所述多个预设关联数据库中的关联分数，确定所述疾病描述
15 词条分别与所述多个基因的关联程度。

- 8、根据权利要求 7 所述的装置，其特征在于，所述关联分数确定单元，
具体用于对多个基因中的任一基因：分别确定所述疾病描述词条与该基因对应
的基因标识在该预设关联数据库中的目标关联记录的记录数据，所述目标关联
记录的记录数据包括：同时包括该基因的基因标识和所述疾病描述词条的关联
20 记录的第一数量、包括所述疾病描述词条的关联记录的第二数量、包括该基因
的基因标识的关联记录的第三数量以及该预设关联数据库中的关联记录的总
数量。

- 9、根据权利要求 8 所述的装置，其特征在于，所述预设词条-基因关联矩
阵为：

$$25 \quad M = \frac{T_1}{T_2} \times \log \left(\frac{T_s}{T_3 + 1} \right)$$

其中， M 为在该预设信息-基因数据库中所述疾病描述词条与该基因对应的
基因标识的关联分数； T_1 为所述第一数量； T_2 为所述第二数量； T_3 为所述第

三数量； T_s 为所述总数量。

10、根据权利要求 7 所述的装置，其特征在于，所述关联程度确定单元包括关联值获得子单元和关联程度确定子单元，

5 所述关联值获得子单元，用于对任一基因：将所述疾病描述词条与该基因对应的基因标识在所述多个预设信息-基因关联数据库中分别对应的关联分数进行加权处理，获得所述疾病描述词条与该基因的关联值；

所述关联程度确定子单元，用于根据所述关联值，确定所述疾病描述词条与该基因的关联程度。

10 11、根据权利要求 10 所述的装置，其特征在于，所述关联程度确定子单元具体用于根据

$$L = \frac{1}{(1 + e^{-z})}$$

确定所述疾病描述词条与该基因的关联程度，其中， L 为所述疾病描述词条与该基因的关联程度； e 为自然常数； z 为所述关联值。

15 12、根据权利要求 7 所述的装置，其特征在于，还包括：预测关联分数确定单元、预测关联程度确定单元、误差确定单元、误差比较单元、词条-基因关联矩阵确定单元以及矩阵参数调整单元，

20 所述预测关联分数确定单元，用于将作为训练信息的多个疾病描述词条与所述多个基因对应的基因标识在所述多个预设信息-基因关联数据库中的目标关联记录的记录数据输入至当前的词条-基因关联训练矩阵中，确定所述训练信息与所述多个基因在各预设信息-基因数据库下分别对应的预测关联分数；

所述预测关联程度确定单元，用于根据确定的各所述预测关联分数，分别确定所述多个疾病描述词条与所述多个基因的预测关联程度；

所述误差确定单元，用于确定所述多个疾病描述词条与所述多个基因的真实关联程度与所述预测关联程度的误差；

25 所述误差比较单元，用于判断所述误差是否处于预设误差阈值以下，如果是，则触发所述词条-基因关联矩阵确定单元，如果不是，则触发所述矩阵参数调整单元；

所述词条-基因关联矩阵确定单元，用于确定所述当前的词条-基因关联训练矩阵为所述词条-基因关联矩阵；

- 所述矩阵参数调整单元，用于通过梯度下降算法调整所述当前的词条-基因关联训练矩阵中的矩阵参数，将调整后的词条-基因关联训练矩阵作为当前的
- 5 的词条-基因关联训练矩阵，触发所述预测关联分数确定单元。

13、一种存储介质，其特征在于，所述存储介质上存储有程序，所述程序被处理器执行时实现权利要求 1 至 6 中任一项所述的基因关联程度确定方法。

14、一种处理器，其特征在于，所述处理器用于运行程序，其中，所述程序运行时执行权利要求 1 至 6 中任一项所述的基因关联程度确定方法。

- 10 15、一种电子设备，其特征在于，所述电子设备包括至少一个处理器、以及与所述处理器连接的至少一个存储器、总线；其中，所述处理器、所述存储器通过所述总线完成相互间的通信；所述处理器用于调用所述存储器中的程序指令，以执行权利要求 1 至 6 中任一项所述的基因关联程度确定方法。

- 16、一种计算机程序产品，其特征在于，当在电子设备上执行时，适于执行初始化权利要求 1 至 6 中任一项所述的基因关联程度确定方法步骤的程序。
- 15