(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 111436212 A
(43)申请公布日 2020.07.21

(21)申请号 201980000228.1

(22)申请日 2019.01.15

(30)优先权数据
201821042893 2018.11.14 IN

(85)PCT国际申请进入国家阶段日
2019.03.07

(86)PCT国际申请的申请数据
PCT/IB2019/050317 2019.01.15

(87)PCT国际申请的公布数据
WO2020/099941 EN 2020.05.22

(71)申请人 库雷人工智能科技私人有限公司
地址 印度孟买

(72)发明人 普雷塔姆·普萨
马努基·塔德帕里
巴尔加瓦·雷迪 塔伦·尼姆马达
普贾·拉奥 普拉桑特·瓦瑞尔

(74)专利代理机构 北京安信方达知识产权代理有限公司 11262
代理人 陆建萍 杨明钊

(51)Int.Cl.
G06K 9/00(2006.01)
A61B 5/05(2006.01)

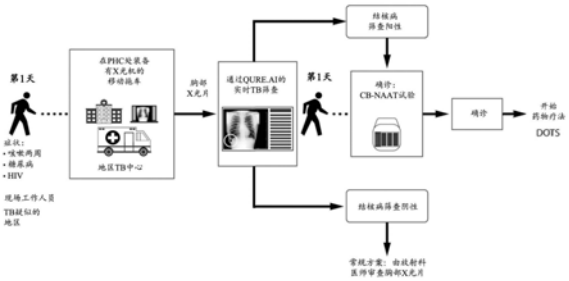
权利要求书1页 说明书12页 附图5页

(54)发明名称

用于医学成像评估的深度学习的应用

(57)摘要

本公开总体上涉及用于处理从成像或其他诊断和评估医疗流程中获得的电子数据的方法和系统。某些实施例涉及用于深度学习算法的开发的方法，该算法对成像和其他医学数据中的特定特征和状况执行机器识别。另一实施例提供了被配置为通过深度学习算法来检测并定位医学成像扫描上的医学异常的系统。



CN 111436212 A

1.一种用于开发深度学习系统以检测并定位胸部X射线扫描上的医学异常的方法,包括:

选择医学成像扫描,并且使用自然语言处理(NLP)算法来提取医学异常;

通过调整大小和标签特定的数据扩增来预处理所选择的医学成像扫描;

用所述选择的医学成像扫描来训练深度学习算法;

通过组合多个模型的预测来预测特定类型的医学异常的存在/不存在;以及

生成对应于所述医学异常的识别水平的分数,并且输出表示所述医学异常的精确位置和范围的边界框。

2.根据权利要求1所述的方法,其中,所述深度学习算法包括卷积神经网络架构。

3.根据权利要求2所述的方法,其中,通过预先训练将胸部X光片与其他身体部位的X光片分离的任务来修改所述架构。

4.根据权利要求1所述的方法,其中,使用各种启发法来选择所述模型。

5.根据权利要求1所述的方法,其中,所述医学异常包括钝化的CP角、钙化、心脏肥大、空洞、实变、纤维化、肺门增大、混浊和胸腔积液。

6.根据权利要求1所述的方法,其中,通过与放射科医师报告进行比较来验证用于检测所述医学异常的所述深度学习算法的准确性。

7.一种被配置为使用深度学习算法来检测并定位胸部X射线扫描上的医学异常的系统,其中,通过以下步骤来开发所述深度学习算法:

选择医学成像扫描,并且使用自然语言处理(NLP)算法来提取医学异常;

通过调整大小和标签特定的数据扩增来预处理所选择的医学成像扫描;

用所述选择的医学成像扫描来训练深度学习算法;

通过组合多个模型的预测来预测特定类型的医学异常的存在/不存在;以及

生成对应于所述医学异常的识别水平的分数,并且输出表示所述医学异常的精确位置和范围的边界框。

8.根据权利要求7所述的系统,其中,所述深度学习算法包括卷积神经网络架构。

9.根据权利要求8所述的系统,其中,通过预先训练将胸部X光片与其他身体部位的X光片分离的任务来修改所述架构。

10.根据权利要求7所述的系统,其中,使用各种启发法来选择所述模型。

11.根据权利要求7所述的系统,其中,所述医学异常包括钝化的CP角、钙化、心脏肥大、空洞、实变、纤维化、肺门增大、混浊和胸腔积液。

12.根据权利要求7所述的系统,其中,通过与放射科医师报告进行比较来验证用于检测所述医学异常的所述深度学习算法的准确性。

13.根据权利要求7所述的系统,其中,所述算法对于异常扫描的检测实现了0.93±0.01的AUC,以及对于钝化的肋膈角、钙化、心脏肥大、空洞、实变、纤维化、肺门增大、混浊和胸腔积液的检测分别实现了0.94±0.02、0.88±0.03、0.97±0.02、0.93±0.07、0.93±0.04、0.88±0.05、0.89±0.05、0.93±0.02、0.98±0.02的AUC。

用于医学成像评估的深度学习的应用

[0001] 相关申请

[0002] 本申请要求享有于2018年11月14日提交的第201821042893号印度专利申请的优先权权益,通过引用将其全部并入本文用于所有目的。

技术领域

[0003] 本公开总体上涉及用于处理从成像或其他诊断和评估医疗程序中获得的电子数据的方法和系统。一些实施例涉及用于深度学习算法的开发的方法,该算法对成像和其他医学数据中的特定特征和状况执行机器识别。

背景技术

[0004] 诸如计算机断层摄影(CT)和X射线成像的医学成像技术广泛用于诊断、临床研究和治疗计划。存在对提高医学成像评估的效率、准确性和成本效益的自动化方法的新兴需求。

[0005] 胸部X光片在最常见的放射学诊断试验当中,每年全球都会进行数百万次扫描。虽然经常进行试验,但是读取胸部X光片成为更复杂的放射学任务,并且众所周知是高度主观的,其中根据评阅者(reader)的经验水平、检测到的异常和临床环境,评阅者之间的一致性从0.2的kappa值到0.77的kappa值变化。

[0006] 由于他们的可支付力(affordability),世界各地,包括在很少有或没有放射科医师的地区都使用胸部X光片。在世界许多地方,数字胸部X光片机的可用性(availability)比经过高度训练以执行这项复杂的任务的临床医生的可用性增长地更快。如果自动检测可以在资源匮乏环境中作为疾病筛查工具被应用,那么对全球人口健康结果的益处可以很有意义。胸部X光片的这种使用的一个示例是在结核病筛查中,其中专家评阅者所掌握的胸部X光片对结核病的早期检测比临床症状更敏感。

[0007] 在过去几年中,人们对使用深度学习算法来帮助医学图像上的异常检测的兴趣越来越大。这是机器解释自然图像和检测其中的对象的能力快速增长的自然结果。特别是在胸部X光片方面,已经有一系列研究描述了使用深度学习算法来检测各种异常(Shin等人,Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,第2497-2506页,2016;Rajpurkar等人,arXiv preprint arXiv:1711.05225,2017;Li等人,arXiv preprint arXiv:1711.06373,2017)。这些中的大部分受到缺乏大型高质量数据集的可用性的限制,其中发表的著作(work)描述的最大的是对112,120张X光片进行训练的算法,相对少数考虑到大多数胸部X光片是正常的(normal),并且异常的(abnormal)X光片不太常见,特定的异常更少。

[0008] 发明概述

[0009] 本公开描述了全自动深度学习系统的开发和临床验证,该系统被训练以从医学成像扫描检测并定位异常。

[0010] 特定实施例提供了深度学习系统的训练和临床验证,以检测并定位胸部X光片异

常。该系统已经在120万张X光片上进行了训练，并在包含2000项研究的独立数据集上对照由3名放射科医师组成的小组(panel)的多数投票对其进行了试验。胸部X光片上异常范围从非常小的病变(lesion)到覆盖肺野(field)大部分的弥漫性(diffuse)异常。最佳深度学习算法架构基于检测到的异常而不同；因此，提供了针对每个异常使用具有最大化AUC的单独算法的系统。

[0011] 具体地，实施例提供了用于开发深度学习系统以检测并定位胸部X射线扫描上的医学异常的方法，包括：

[0012] 选择医学成像扫描，并且使用自然语言处理(NLP)算法来提取医学异常；

[0013] 通过调整大小和标签特定的数据扩增(augmentation)来预处理所选择的医学成像扫描；

[0014] 用选择的医学成像扫描来训练包括卷积神经网络架构的深度学习算法，其中，通过预先训练将胸部X射线扫描与其他身体部位的X射线扫描分离的任务来修改架构；

[0015] 通过组合多个模型的预测来预测特定类型的医学异常的存在/不存在，其中，使用各种启发法(heuristic)来选择模型；

[0016] 生成对应于医学异常的识别水平的分数，并且输出表示医学异常的精确位置和范围的边界框；以及

[0017] 通过与放射科医师报告进行比较来验证用于检测医学异常的深度学习算法的准确性。

[0018] 根据实施例，所述医学成像扫描包括但不限于CT、X射线、核磁共振成像(MRI)和超声流程。对于胸部X射线扫描，所述医学异常包括但不限于钝化的CP角(blunted CP angle)、钙化(calcification)、心脏肥大(cardiomegaly)、空洞(cavity)、实变(consolidation)、纤维化(fibrosis)、肺门增大(hilar enlargement)、混浊(opacity)和胸腔积液(pleural effusion)。

[0019] 另一实施例提供了被配置为通过深度学习算法来检测并定位胸部X射线扫描上的医学异常的系统，其中，通过以下步骤来开发深度学习算法：

[0020] 选择医学成像扫描，并且使用自然语言处理(NLP)算法来提取医学异常；

[0021] 通过调整大小和标签特定的数据扩增来预处理所选择的医学成像扫描；

[0022] 用选择的医学成像扫描来训练包括卷积神经网络架构的深度学习算法，其中，通过预先训练将胸部X射线扫描与其他身体部位的X射线扫描分离的任务来修改架构；

[0023] 通过组合多个模型的预测来预测特定类型的医学异常的存在/不存在，其中，使用各种启发法来选择模型；

[0024] 生成对应于医学异常的识别水平的分数，并且输出表示医学异常的精确位置和范围的边界框；以及

[0025] 通过与放射科医师报告进行比较来验证用于检测医学异常的深度学习算法的准确性。

[0026] 此外，系统被配置为通过深度学习算法来检测并定位胸部X射线扫描上的医学异常的系统，其中，算法实现了对于异常扫描的检测的AUC为0.93±0.01，以及对于钝化的肋膈角、钙化、心脏肥大、空洞、实变、纤维化、肺门增大、混浊和胸腔积液的检测的AUC分别为0.94±0.02、0.88±0.03、0.97±0.02、0.93±0.07、0.93±0.04、0.88±0.05、0.89±0.05、

0.93±0.02、0.98±0.02。

[0027] 发明的有益效果

[0028] 以前发表的关于用于胸部X光片异常检测的深度学习的著作没有区分“疾病”的诊断和“异常调查结果(findings)”的检测。本发明的重点是检测异常调查结果或者X光片上的异常,这种异常可以由专家在没有临床历史的任何先验知识(prior knowledge)的情况下可视地检测出来。这允许将系统应用于具有不同疾病患病率(prevalence)模式的地理环境和不同疾病表现。

[0029] 附图简述

[0030] 图1对于深度学习系统以检测胸部X光片异常的放射科医师验证的研究设计。

[0031] 图2对于所有异常的AUC曲线相对于3名放射科医师多数,其中标记了评阅者性能。

[0032] 图3使用qXR在地区TB中心进行试点(pilot)的TB检测的推荐工作流。

[0033] 详细描述

[0034] 应当理解,本发明不限于本文描述的特定的方法论、协议和系统等,且因此可以变化。本文使用的术语仅为了描述特定实施例的目的,而不旨在限制本发明的范围,本发明的范围仅由权利要求限定。

[0035] 如说明书和所附权利要求中所使用的,除非有相反的指定,否则以下术语具有以下所指示的含义。

[0036] “架构”是指描述计算机系统的功能、组织和实现的一组规则和方法。

[0037] “卷积神经网络(CNN)”是指一类深度、前馈人工神经网络,最常用于分析视觉影像。CNN使用多层感知器的变体,这些感知器被设计为需要最少的预处理。CNN由输入和输出层以及多个隐藏层组成。CNN的隐藏层通常由卷积层、池化(pooling)层、全连接层和标准化层组成。卷积层对输入应用卷积运算,将结果传递给下一层。局部或全局池化层将一层处的神经元簇(cluster)的输出合并到下一层中的单个神经元中。全连接层将一层中的每个神经元连接到另一层中的每个神经元。与其他图像分类算法相比,CNN使用相对较少的预处理。这意味着网络学习传统算法中手工设计(hand-engineered)的过滤器。在特征设计中,独立于先前知识和人类努力是主要的优势。

[0038] “启发法”是指被设计用于在经典方法太慢时更快地解决问题,或者用于在经典方法找不到任何精确解(exact solution)时找到近似解的技术。这是通过交易最优性、完整性、准确性或速度精度来实现的。从某种意义上来说,这可以被认为是捷径。启发式函数(也简称为启发法)是基于可用信息在每个分支步骤中对搜索算法中的备选方案进行排序以决定跟随哪个分支的函数。启发法的目标是在合理的时间范围内产生足以解决手头问题的解。该解可能不是对该问题的所有解中最好的,或者它可以仅仅近似精确解。

[0039] “自然语言处理(NLP)”是指计算机以智能且有用的方式来分析、理解人类语言并从人类语言获得含义的方法。通过利用NLP,开发人员可以组织并构造知识来执行任务,诸如自动摘要、翻译命名实体识别、关系提取、情感分析、语音识别和主题分割。

[0040] 本公开示出了能够在数据驱动的图像评估工作流中集成并使用机器学习分析的各种技术和配置。例如,可以对被产生作为医学成像研究的一部分的医学成像过程数据执行机器学习分析(诸如某些医学状况的图像检测的经训练的模型)。医学成像过程数据可以包括由成像模态捕获的图像数据,以及命令(order)数据(诸如指示对放射线图像读取的请

求的数据),每个都是为了便于医学成像评估而产生的(诸如由放射科医师执行的放射学读取或者由另一合格的医学专业人员执行的诊断评估)。

[0041] 例如,机器学习分析可以接收并处理来自医学成像过程数据的图像,以识别经训练的结构、状况以及特定研究的图像内的状况。机器学习分析可以导致自动检测、指示或确认图像内的某些医疗状况,例如紧急或生命攸关的医疗状况、临床严重异常和其他关键调查结果。基于机器学习分析的结果,对于图像的医学评估和相关联的成像过程可以被设置优先级,或者以其他方式被改变或被修改。此外,医学状况的检测可以用于帮助将医学成像数据分配给特定评估者、对于医学成像数据的评估过程,或在医学成像评估之前或与医学成像评估同时执行其他动作(或生成诸如来自这样的医学成像评估的报告的数据项)。

[0042] 如本文进一步讨论的,机器学习分析可以代表任何数量的机器学习算法和经训练的模型来被提供,包括但不限于已经被训练来执行图像识别任务的深度学习模型(也称为深度机器学习或分层模型),特别是对于人体解剖和解剖表示的医学图像上的某些类型的医学状况。如本文所使用的,术语“机器学习”用于指各类人工智能算法和能够执行经训练的结构的机器驱动(例如,计算机辅助)识别的算法驱动方法,其中术语“深度学习”是指使用多层表示和抽象的这种机器学习算法的多层操作。然而,明显的是,在当前描述的医学成像评估中应用、使用和配置的机器学习算法的角色可以由任何数量的其他基于算法的方法(包括人工神经网络的变体、有学习能力的算法、可训练的对象分类和其他人工智能处理技术)来补充或替代。

[0043] 在下面的一些示例中,参考放射学医学成像过程(例如,计算机断层摄影(CT)、核磁共振成像(MRI)、超声波和X射线过程等)和由这样的成像过程产生的图像的诊断评估,这样的成像过程由经许可并认证的放射科医师通过图像评估(例如,放射学读取)来执行。应当理解,当前描述的技术和系统的适用性将扩展到由各种医学过程和专业(包括不涉及传统放射学成像模态的那些)产生的各种各样的成像数据(和其他数据表示)。这样的专业包括但不限于病理学、医学摄影、诸如脑电图学(EEG)和心电图学(EKG)过程的医学数据测量、心脏病学数据、神经科学数据、临床前成像以及与远程医疗、远程病理学(telepathology)、远程诊断相关的其他数据收集过程,以及医学过程和医学科学的其他应用。因此,本文描述的数据识别和工作流修改技术的性能可以应用于各种医学图像数据类型、设置和用例(use case),包括捕获的静态图像和多图像(例如视频)表示。

[0044] 以下描述和附图充分说明了具体实施例,以使本领域技术人员能够实施它们。其他实施例可以结合结构、逻辑、电气、过程和其他变化。一些实施例的部分及特征可以被包括在其它实施例中或是代替其他实施例的部分及特征。

[0045] 示例

[0046] 示例1深度学习系统以检测胸部X光片异常的放射科医师验证

[0047] 1.1.方法

[0048] 1.1.1算法开发

[0049] 1,200,000张X光片及其相应的放射学报告被用于训练卷积神经网络(CNN)以识别异常。开发了自然语言处理算法来解析(parse)非结构化放射学报告,并提取关于胸部X光片中的异常的存在的信息。在训练CNN时,这些提取的调查结果被用作标注(label)。训练单个网络以识别正常的X光片,以及胸部X光片调查结果“钝化的CP角”、“钙化”、“心脏肥大”、

“空洞”、“实变”、“纤维化”、“肺门增大”、“混浊”和“胸腔积液”。表1列出了从报告中提取放射学调查结果时使用的定义。这些调查结果被称为标签(tag)。标签提取准确性相对于一组手动提取异常的报告进行测量。表2a报告了标签提取准确性。

[0050] 表1.异常定义

[0051]

调查结果	来自放射学报告的标签提取的定义	放射科医师审查的定义
正常	“未检测到异常”或“正常”	正常的X光片
钝化的CP角	钝化的CP角	CP角度钝化或模糊。可能是由于胸腔积液或胸膜增厚（pleural thickening）
钙化	钙化	X光片上的所有钙化包括但不限于主动脉弓（aortic arch）钙化、肋骨（rib）钙化、钙化肺密度和微钙化（microcalcifications）
心脏肥大	心脏肥大	心胸比率>0.5
空洞	肺空洞	肺空洞
实变	实变、肺炎（pneumonia）或充气支气管造影（air-bronchogram）	肺实变
纤维化	纤维化	任何肺纤维化的证据，包括间质（interstitial）纤维化、纤维空洞性病变（fibrocavitary lesion）
肺门突起(Hilar prominence)	肺门增大，突出门（prominent hilum）或肺门淋巴结病（hilar lymphadenopathy）	增大的或突起的门或包括肺门淋巴结病
混浊	任何肺野混浊（optacity）或混浊（optacities）、阴影或密度，包括但不限于浸润（infiltrate）、实变、肿块（mass）、结节（nodule）、肺钙化和纤维化	任何肺野混浊或多重混浊（multiple opacities），包括但不限于浸润、实变、肿块、结节、肺钙化和纤维化。胸腔异常不被包括在该标签中
胸腔积液	胸腔积液	胸腔积液

[0052]

[0053] 表2.标签提取准确性。

[0054]

调查结果	#阳性	敏感性（Sensitivity）（95% CI）	特异性（Specificity）（95% CI）
正常（未检测到异常）	105	0.9429 (0.8798-0.9787)	1.0000 (0.9959-1.0000)
钝化的 CP 角	146	0.9795 (0.9411-0.9957)	0.9824 (0.9712-0.9901)
钙化	116	1.0000 (0.9687-1.0000)	0.9660 (0.9519-0.9770)
心脏肥大	125	0.9920 (0.9562-0.9998)	0.9760 (0.9635-0.9851)
空洞	30	1.0000 (0.8843-1.0000)	0.9856 (0.9759-0.9921)
实变	161	0.9876 (0.9558-0.9985)	0.9761 (0.9634-0.9854)
纤维化	124	0.9839 (0.9430-0.9980)	0.9931 (0.9851-0.9975)
肺门增大	289	0.9689 (0.9417-0.9857)	0.9732 (0.9585-0.9838)
混浊	612	0.9608 (0.9422-0.9747)	0.9251 (0.8942-0.9492)
胸腔积液	246	0.9309 (0.8917-0.9592)	0.9602 (0.9436-0.9730)
共计（所有调查结果）	1954	0.9672 (0.9584-0.9747)	0.9771 (0.9736-0.9803)

[0055] (a)标签提取准确性

[0056]

调查结果	放射科医师 1 和 2 一致%	Cohen's κ	所有读数 Fleiss' κ
正常（未检测到异常）	85.00	0.6049	0.5618
钝化的 CP 角	83.58	0.2968	0.3054
钙化	83.00	0.5803	0.5821
心脏肥大	91.60	0.5333	0.5284
空洞	97.50	0.3824	0.4047
实变	88.28	0.3529	0.3397
纤维化	89.40	0.3781	0.3495
肺门增大	89.38	0.2630	0.2101
混浊	70.70	0.2306	0.1733
胸腔积液	90.69	0.5341	0.5305

[0057] (b)放射科医师之间的一致性。Fleiss'sκ是在两个放射科医师的读数和用NLP从原始报告提取的实际真相(ground truth)之间计算的。

[0058] 1.1.1.1预处理和数据扩增

[0059] 所有X光片都被调整大小到标准大小，并且应用了一组标准归一化来减少源相关的变化。调整大小的操作包括对原始图像进行下采样。虽然下采样意味着潜在的有价值的信息的丢失，但是它通过克服维数灾难来帮助模型更好地训练，并且由AI的当前状态进行保证。此外，图像的大小会显著影响推理的速度。

[0060] 鉴于X光片不如其他医学成像模态(modality)(如CT和MRI)那样标准化，在训练的同时使用了大量标签特定的数据扩增，其旨在使模型对被用于获取、曝光和噪声的制造商/模型中的可变性具有鲁棒性。

[0061] 1.1.1.2架构和训练

[0062] 检测个体异常的系统中的基本块是DenseNet(Huang等人，Proceedings of the

IEEE conference on computer vision and pattern recognition,第1卷,第3页,2017)或ResNet(He等人,Proceedings of the IEEE conference on computer vision and pattern recognition,第770-778页,2016)的修改版本。构建个体异常检测系统的所有分类模型都是针对将胸部X光片与其他身体部位的X光片分离的任务进行预训练的,而不是流行的ImageNet预训练。该步骤旨在利用由所有X光片组成的超级集合。在与ImageNet预训练相比时,观察到改进的模型收敛(convergence)和在泛化(generalization)性能上的增加的增益。

[0063] 1.1.1.3模型集成的使用

[0064] 集成(ensembling)是通过组合由模型的集合产生的预测来提高泛化性能的简单方法。使用不同的初始条件、架构和采样(训练时使用的异常的分布)来创建多个模型以检测特定异常。使用各种启发法来选择这些模型的子集,并且使用多数集成方案来组合这些选择的模型的预测,以做出关于特定异常的存在/不存在的决定。

[0065] 1.2研究设计

[0066] 1.2.1采样大小和X光片选择

[0067] 对于研究使用了来自南印度的两个哥伦比亚亚洲医院中心的门诊和住院X光片的组合。在该研究中使用之前,所有的数据都被去识别(de-identified)。放射科医师验证分两个阶段进行,每个阶段使用1000张X光片单独集合。章节1.4.1中解释了样本大小的计算。来自两个中心的所有胸部X光片都按如下方式进行过滤。

[0068] 1.选择所有PA和AP位(view)胸部X光片,其中放射科医师报告是可用的。

[0069] 2.在这个集合中,排除了来自儿科患者(年龄≤14岁)的X光片,排除了以仰卧位拍摄的X光片(床边/便携式X光片)。

[0070] 从该池中随机选择了1000张X光片的集合(集合1),没有包含感兴趣的异常的X光片的富集(enrichment)。从相同的池中采样1000张X光片的第二集合(集合2)(不包括先前选择的1000张X光片),使得它们包括每个异常的至少80个示例(可用的情况下),其余的X光片是随机选择的。解析X光片放射学报告的自然语言处理工具被用于实现上面列出的排除,并且自动化对于阶段2的采样。图1中示出了研究设计。

[0071] 1.2.2包含和排除准则

[0072] 研究中包括了以站立姿势拍摄的非卧床(ambulatory)成人患者的PA和AP位胸部X光片。排除了来自14岁以下患者的X光片、来自床边或便携式X光机的以仰卧位拍摄的X光片。因此,数据集不包含任何带有可见静脉置(intravenous lines)、管状物(tube)或导管(catheter)或ECG导联(lead)的X光片,或带有诸如起搏器的植入医疗设备的患者的X光片。

[0073] 1.2.3异常定义

[0074] 为了确保审查(review)放射科医师和算法在相同的参照系下操作以用于定义异常,以下工作定义用于“正常”和每个异常。在算法开发阶段期间,这些定义用于从报告中提取标签,并且在验证阶段期间,由放射科医师使用这些定义。

[0075] 1.3放射科医师验证和黄金标准创建

[0076] 1.3.1从报告中提取异常

[0077] 一种结合自然语言处理算法的定制的基于词典的方法被用于从对应于X光片的原始放射科医师报告中提取相关调查结果。自然语言处理算法的准确性被量化,用于从由

2000次扫描组成的独立数据集上的非结构化放射学报告中提取胸部X光片异常，与对于每次扫描只有单个专家评阅者的意见进行对比。专家们被提供表1中的标签定义、原始报告以及胸部X光片，并且无视于算法输出。

[0078] 1.3.2对于从图像中异常检测的深度学习算法准确性的验证

[0079] X光片被随机分配给6名经委员会认证的放射科医师(3对)，他们具有3至15年的放射学专业知识，使得每一张X光片都由2名放射科医师独立评阅，这2名放射科医师都没有最初地报告过该X光片。第三读数以附带X光片的原始放射学报告的形式可用。放射科医师验证使用了定制验证门户(portal)。DICOM文件有完整的原始分辨率可供放射科医师在决定是否存在异常时使用。放射科医师能够针对每个异常标记“存在”或“不存在”，也可以使用钢笔工具来标记出受影响的区域。附加的文本域可用于反馈。

[0080] 对于该研究的黄金标准(gold standard)是3名放射科医师读数之间对于每个异常的存在或不存在的大多数意见(2名放射科医师使用验证门户用于回顾性(retrospective)评估以及1份原始放射学报告)。相对于该黄金标准报告了检测异常X光片和每个单独异常的算法准确性。算法输出在两个数据集上生成，并且放置在锁定的数据库中，直到放射科医师验证完成。放射科医师在针对验证研究审查X光片时，无视于原始X光片报告和算法输出。

[0081] 1.4统计分析

[0082] 1.4.1样本大小计算

[0083] 以95％的置信区间精确地计算敏感性和特异性将需要从人群中随机采样的未富集数据集中的大约20,000个的样本大小。使用了章节1.2.1中详述的富集策略。以90％准确性和95％置信区间估计80％敏感性所需的样本为951。因此，2000张X光片足以评估X光片作为正常和异常的分类。由于在初步诊断调查中，假阴性被认为比假阳性更严重，因此对于该样本大小计算考虑了敏感性。

[0084] 在典型的门诊环境中，异常的患病率低于5％。951张X光片的样本大小可以在95％置信区间上给出3％的精确度来估计80％的特异性。鉴于对于以上列出的每种具体状况的患病率低于10％，每种状况下至少77个真实病例将以10％精确度和95％的置信区间给出80％的敏感性。因此，对于集合2，每种状况最少使用80个真实病例。

[0085] 1.4.2算法性能度量相对于放射科医师

[0086] 当计算对于特定异常的AUC时，使用集合1的整体，以及集合2中对于该异常为阳性的所有X光片。

[0087] 1.使用足够大量的“异常阳性”病例进行敏感性计算。

[0088] 2.保持正常的比率:异常接近自然分布。与使用所组合的两个集合比较时，这使得对特异性的评估更加公平。

[0089] 实际上，这相当于进行独立的研究，其单独地评估每个异常的检测的准确性。然而，在计算对于异常X光片的检测的AUC时，使用了整个数据集(n＝2000)。

[0090] 使用由Hanley和McNeil描述的“基于分布的”方法(Radiology,143(1):29-36,1982)来计算AUC置信区间。使用一致的百分比和Cohen's kappaκ统计(Fam Med,37(5):360-363,2005)来测量成对评阅者之间对每项调查结果的一致性。此外，使用Fleiss'kappaκ统计(Psychological bulletin,76(5):378,1971)来测量所有三名评阅者(2名放射科医

师和原始报告)之间对每项调查结果的一致性。因为评阅者被分组成3对,所以评阅者之间一致性的每个报告的测量结果是3个估计值的平均值。

[0091] 1.3结果

[0092] 表3汇总了基本人口统计和包含每个异常的扫描的次数。2000张X光片中有658张是异常的,其中最常见的异常是“混浊”、“心脏肥大”和“钙化”。没有足够的包含空洞的扫描来可信地计算深度学习系统在识别该异常上的准确性。

[0093] 表3.研究人群的人口统计。

[0094]

特性	集合 1 (n=1000)	集合 2 (n=1000)	组合的 (n=2000)
患者人口统计			
年龄			
年龄已知的扫描的次数	803	1000	1803
平均值	48.04	50.61	49.32
标准偏差	18.69	17.93	18.36
范围	16 - 95	16 - 100	16 - 100
女性的数量/性别已知的扫描的次数(百分比)	324/803 (40.3%)	265/1000 (26.5%)	589/1803 (32.6%)
患病率			
扫描的次数(百分比)			
未检测到异常(正常)	440	218	658
钝化的 CP 角	35	121	156
钙化	57	244	301
心肥大	61	116	177
空洞	1	15	16
实变	13	92	105
纤维化	13	97	110
肺门增大	15	49	64
混浊	104	341	445
胸腔积液	36	122	158

[0095] 实现报告解析的高准确性有助于能够使用大量X光片来训练深度学习算法。表2a汇总了来自放射学报告的异常提取准确性相对于由单个评阅者手动提取。算法能够相对于专家评阅者检测正常的X光片报告,具有0.94的敏感性以及1的特异性。对于报告中的个体异常的检测,敏感性从胸腔积液的0.93到钙化和空洞的1变化;特异性从混浊的0.92到纤维化的0.99变化。

[0096] 1.3.1放射科医师之间的一致性

[0097] 表2b描述了评阅者之间的一致性。异常X光片的检测的一致性最高(评阅者之间的一致性为85%,Cohen's Kappa为0.6,Fleiss'kappa为0.56)。

[0098] 1.3.2算法准确性相对于3名放射科医师的大多数意见

[0099] 表4列出了在识别12种异常的每一种上的深度学习系统准确性。图2示出了对于在整个数据集(n=2000)上绘制的每个异常的AUC曲线,相对于从6名放射科医师中随机抽取(drawn)的3名放射科医师的大多数意见(2名放射科医师回顾性地重新评阅,加上原始放射学报告)。在每张图上标记了对于6名放射科医师的个体放射科医师敏感性和特异性。在大

多数情况下,个体放射科医师敏感性和特异性略高于AUC曲线(胸腔积液、心脏肥大和混浊除外),并且其中算法性能等同于某些个体的性能(或放射科医师敏感性和特异性)。

[0100] 表4.算法的性能

[0101]

调查结果	AUC (95% CI)	高敏感性操作点		高特异性操作点	
		敏感性	特异性	敏感性	特异性
正常(未检测到异常)	0.9278 (0.9162-0.9394)	0.9100	0.7519	0.8192	0.9086
钝化的CP角	0.9434 (0.9179-0.9690)	0.9487	0.7692	0.8461	0.9100
钙化	0.8812 (0.8550-0.9073)	0.9069	0.6550	0.6511	0.8912
心脏肥大	0.9670 (0.9484-0.9856)	0.9717	0.8636	0.8474	0.9284
空洞	0.9351 (0.8511-1.0000)	0.9375	0.7528	0.8125	0.9850
实变	0.9276 (0.8929-0.9623)	0.9238	0.6970	0.8285	0.9112
纤维化	0.8851 (0.8436-0.9266)	0.9000	0.6641	0.7454	0.8669
肺门增大	0.8908 (0.8379-0.9437)	0.9218	0.7509	0.6718	0.8907
混浊	0.9296 (0.9124-0.9468)	0.9191	0.6991	0.8179	0.9097
胸腔积液	0.9841 (0.9703-0.9979)	0.9810	0.8355	0.8797	0.9660

[0102]

[0103] 另一数据集(Qure90k)是在印度的4个中心随机采样的,这些中心对原始训练数据集没有贡献。使用自然语言处理(NLP)在原始放射科医师报告上建立了对于该数据集的参考标准(实际真相)。表5列出了在识别12种异常的每一种上的深度学习系统准确性。

[0104] 表5.算法在Qure90K数据集上的性能。

[0105]

调查结果	AUC（95% CI）	高敏感性操作点		高特异性操作点	
		敏感性	特异性	敏感性	特异性
正常（未检测到异常）	0.8436 (0.8314-0.8558)	0.9000	0.6419	0.6918	0.9000
钝化的CP角	0.9141 (0.9060-0.9223)	0.9000	0.7406	0.7885	0.9000
钙化	0.8622 (0.8360-0.8883)	0.9000	0.6246	0.6736	0.8815
心脏肥大	0.9369 (0.9312-0.9426)	0.9000	0.8450	0.8439	0.9000
空洞	0.9356 (0.9104-0.9608)	0.9000	0.8268	0.8636	0.9000
实变	0.9180 (0.9092-0.9267)	0.9000	0.8101	0.7873	0.9000
纤维化	0.8916 (0.8783-0.9049)	0.9000	0.6512	0.7302	0.9000
肺门增大	0.8501 (0.8379-0.9437)	0.9000	0.5597	0.6022	0.9000
混浊	0.8948 (0.8898-0.8998)	0.9000	0.7321	0.7800	0.9000
胸腔积液	0.9076 (0.9009-0.9144)	0.9000	0.7433	0.7914	0.9000

[0106] 1.4结论

[0107] 深度学习系统示出了对于异常扫描的检测的AUC为0.93(CI0.92-0.94)，以及对于钝化的肋膈角、钙化、心脏肥大、空洞、实变、纤维化、肺门增大、混浊和胸腔积液的检测的AUC(CI)分别为0.94(0.92-0.97)、0.88(0.85-0.91)、0.97(0.95-0.99)、0.93(0.85-1)、0.93(0.89-0.96)、0.88(0.84-0.93)、0.89(0.84-0.94)、0.93(0.91-0.95)、0.98(0.97-1)。

[0108] 该示例表明，用大量良好标注的数据来训练的深度学习算法可以准确地检测胸部X光片上的异常。随着这些系统变得越来越准确，使用它们来扩大胸部X光片解释的范围并提高其效率的可行性增加了。

[0109] 示例2用于结核病检测的深度学习解决方案qXR

[0110] Qure.ai's qXR被设计用于筛查异常胸部X光片并确定其优先级。算法自动识别15种最常见的胸部X光片异常。这些异常的子集暗示了典型或不典型的肺结核被组合在一起，以在结果内生成“肺结核筛查”算法。结核病筛查算法旨在复制放射科医师或内科医师的胸部X光片的筛查，用于微生物确认前的结核病的异常提示。qXR是第一个通过CE认证的基于AI的胸部X光片解释软件。qXR与供应商中立集成流程(Vendor Neutral Integration Process)集成在一起，并与从任何X光片系统(CR或DR)生成的X光片一起工作。qXR筛查结核病，并且也识别15种其他异常，这样就可以告知患者他们可能患有的non-TB状况。

[0111] qXR与供应商中立档案(VNA)和PACS无缝地集成在一起，不会干扰现有的放射学工作流。它被部署在符合HIPAA的云服务器中，以提供包含以下内容的结果的表格：

[0112] a.对于每张胸部X光片是“正常”或“异常”。

[0113] b.算法检测到的所有异常及其相应的概率分数的列表。

[0114] c.对于每张X光片的“结核病筛查建议”或“结核病筛查阴性”，以

[0115] 及概率分数。

[0116] 使用标准成像方案，qXR自动从客户端PACS/VNA上传感兴趣的X光片(匿名化的)，

并且使用HL7格式回复结果叠加(overlay)和相应的报告，允许用户在审查原始X光片之前访问分析。该标准过程确保了有如两周以上咳嗽、糖尿病和其他免疫缺陷的患者的症状的个体在几秒钟内就能完成他们对结核病的筛查，称为微生物确认(NAAT)，并且如果发现是阳性的，治疗可以在同一天开始。图3示出了对于使用qXR在地区TB中心处进行试点的TB检测的推荐工作流。

[0117] 部署工作将包括通过API与当前X光片系统集成在一起。结果与任何放射学查看器兼容，或者可以托管(hosted)在基于网络的Qure查看器上。自动报告是用与图像数据中特定医学状况的特性的识别水平相对应的分数创建的。表6是自动报告的示例。

[0118] 表6.自动报告的示例。

[0119] 胸部X光片异常检测和分数

[0120]

X光片调查结果	概率	备注
异常	0.94	是
钝化的肋膈角	0.24	否
钙化	0.78	是
心脏肥大	0.11	否
空洞	0.85	是
颈肋	0.42	否
实变	0.92	是
过度膨胀(Hyper Inflation)	0.44	否
纤维化	0.8	是
肺门区域中的突起	0.91	是
混浊	0.95	是
胸腔积液	0.08	否
脊柱侧凸	0.1	否
结核病筛查	0.96	建议

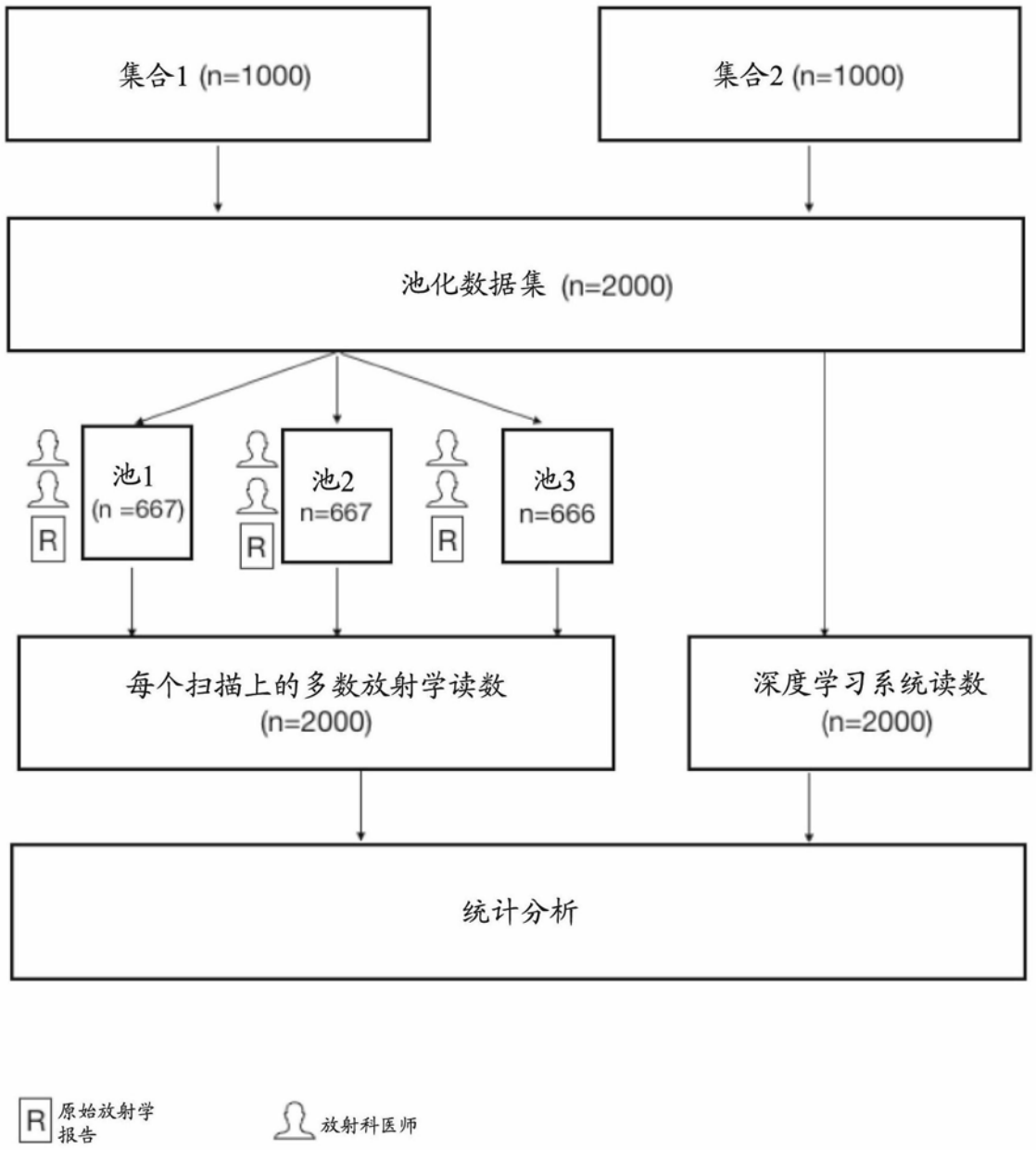

集合1 (n=1000)
集合2 (n=1000)
池化数据集 (n=2000)
池1
(n =667)
池2
n=667
池3
n=666
R
R
R
每个扫描上的多数放射学读数
(n=2000)
深度学习系统读数
(n=2000)
统计分析
R 原始放射学报告
放射科医师

图1

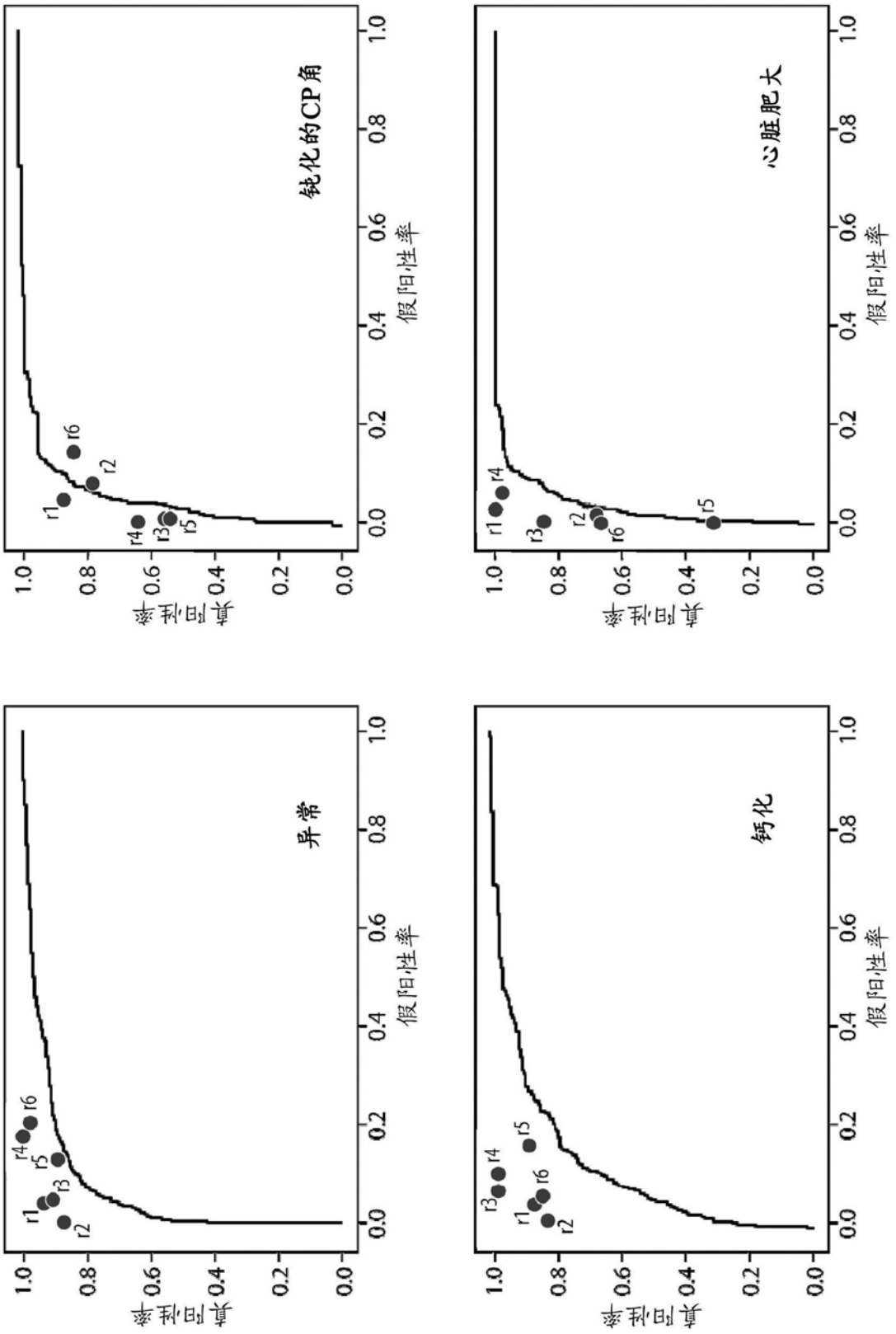


图2

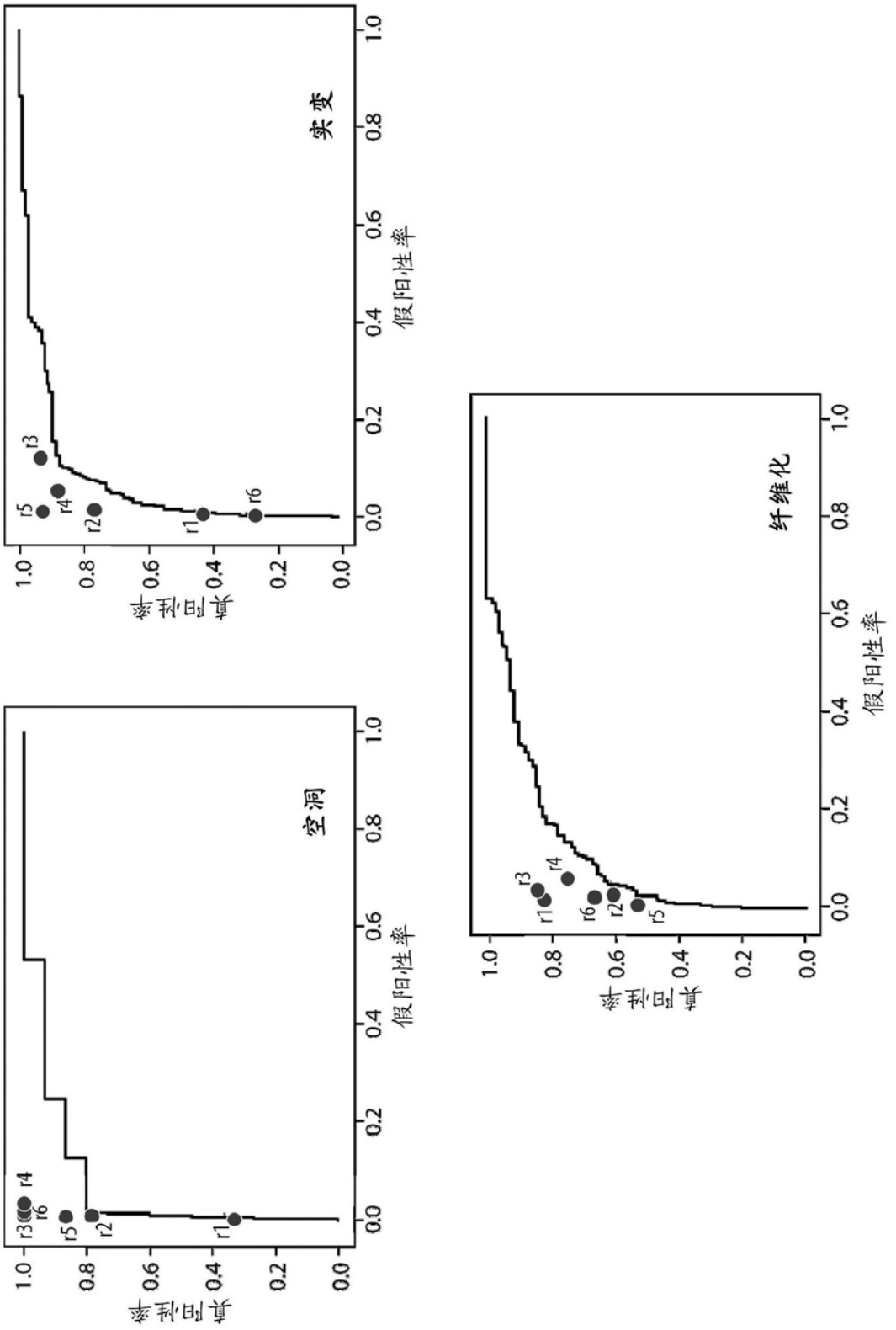


图2继续

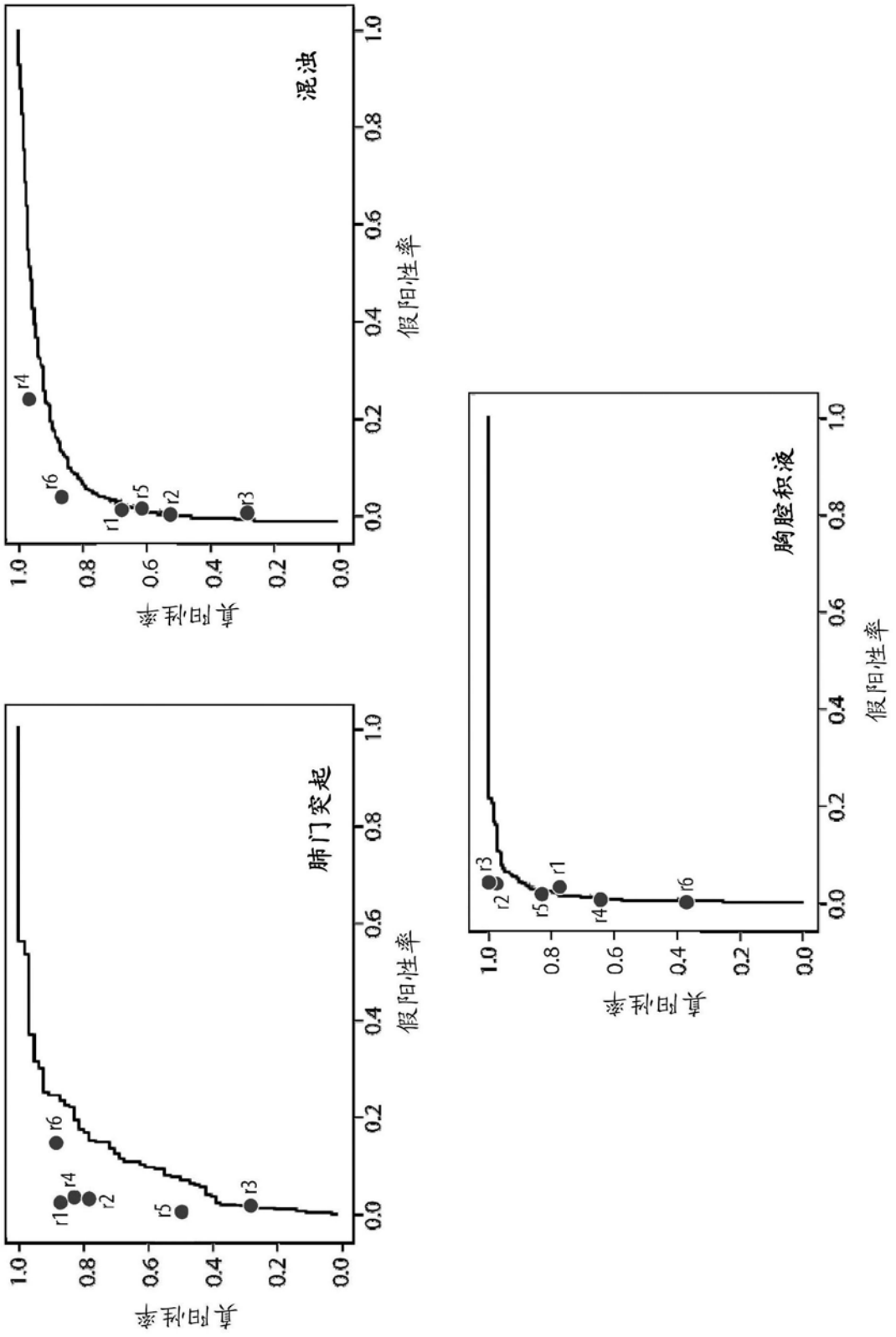


图2继续

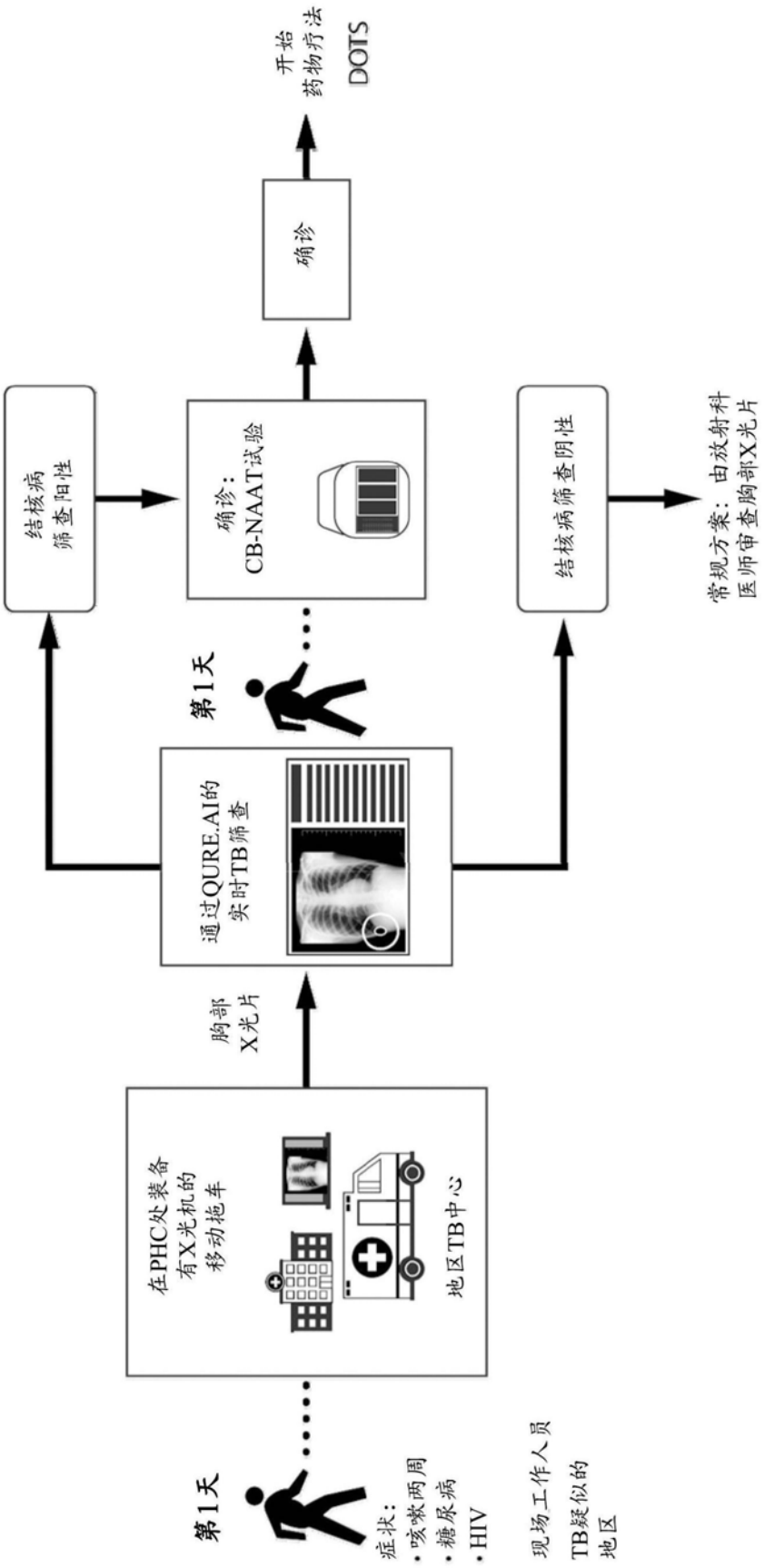

现场工作人员
TB疑似的
地区
症状：
• 咳嗽两周
• 糖尿病
• HIV
第1天
在PHC处装备
有X光机的
移动拖车
地区TB中心
胸部
X光片
通过QURE.AI的
实时TB筛查
结核病
筛查阳性
第1天
确诊：
CB-NAAT试验
确诊
开始
药物疗法
DOTS
结核病筛查阴性
常规方案：由放射科
医师审查胸部X光片

图3