DOI: https://doi.org/10.17816/DD60622

Вариабельность заключений при интерпретации КТ-снимков: один за всех и все за одного

Н.С. Кульберг[1, 2], Р.В. Решетников[1, 3], В.П. Новик[1], А.Б. Елизаров[1], М.А. Гусев[1, 4], В.А. Гомболевский[1], А.В. Владзимирский[1], С.П. Морозов[1]

[1] Научно-практический клинический центр диагностики и телемедицинских технологий Департамента здравоохранения г. Москвы, Москва, Российская Федерация
[2] Федеральный исследовательский центр «Информатика и управление» Российской академии наук, Москва, Российская Федерация
[3] Первый Московский государственный медицинский университет имени И.М. Сеченова (Сеченовский Университет), Москва, Российская Федерация
[4] Московский политехнический университет, Москва, Российская Федерация

АННОТАЦИЯ

Обоснование. Разметка наборов медицинских изображений во многом полагается на субъективную интерпретацию наблюдаемых подозрительных структур. На настоящий момент не существует рекомендованного протокола по определению эталонных данных (ground truth), основанных на врачебных описаниях.

Цель — анализ правильности и согласованности оценок рентгенологов, принимавших участие в подготовке общедоступного набора данных CTLungCa-500; определение взаимосвязи этих показателей с количеством специалистов, проводящих независимую интерпретацию изображений, полученных при компьютерно-томографическом (КТ) исследовании.

Материал и методы. Набор данных, в разметке которого принимали участие 34 рентгенолога, включает 536 КТ-исследований пациентов из группы риска развития рака лёгкого. Каждое КТ-исследование было независимо интерпретировано шестью специалистами, после чего обнаруженные ими подозрительные структуры проходили арбитраж другим экспертом. Для каждого эксперта подсчитывали количество истинно положительных, ложноположительных, истинно отрицательных и ложноотрицательных находок, на основании которых проводили оценку диагностической точности рентгенологов. Для анализа согласованности между заключениями рентгенологов использовали метрику процентного показателя.

Результаты. Увеличение количества специалистов, проводящих независимую интерпретацию КТ-исследований, ведёт к росту правильности их оценок при снижении согласованности. Среди факторов, влияющих на согласованность заключений между парами исследователей, выделяется расхождение мнений по поводу наличия лёгочного очага в конкретном участке КТ-снимка.

Заключение. Увеличение числа независимых первичных интерпретаций способно повысить их комбинированную правильность при условии проведения арбитража, причём квалификация рентгенологов не имеет определяющего значения для качества анализа. Проведение первичной разметки силами четырёх рентгенологов является оптимальным с точки зрения сочетания правильности интерпретации и её стоимости.

Ключевые слова: компьютерная томография; набор данных; эталонные данные; согласованность между заключениями.

Как цитировать

Кульберг Н.С., Решетников Р.В., Новик В.П., Елизаров А.Б., Гусев М.А., Гомболевский В.А., Владзимирский А.В., Морозов С.П. Вариабельность заключений при интерпретации КТ-снимков: один за всех и все за одного // *Digital Diagnostics*. 2021. Т. 2, № 2. С. 105–118. DOI: https://doi.org/10.17816/DD60622

Рукопись получена: 11.02.2021 **Рукопись одобрена:** 07.07.2021 **Опубликована:** 13.07.2021

DOI: https://doi.org/10.17816/DD60622

Inter-observer variability between readers of CT images: all for one and one for all

Nikolas S. Kulberg[1, 2], Roman V. Reshetnikov[1, 3], Vladimir P. Novik[1], Alexey B. Elizarov[1], Maxim A. Gusev[1, 4], Victor A. Gombolevskiy[1], Anton V. Vladzymyrskyy[1], Sergey P. Morozov[1]

[1] Moscow Center for Diagnostics and Telemedicine, Moscow, Russian Federation
[2] Federal Research Center "Computer Science and Control" of Russian Academy of Sciences, Moscow, Russian Federation
[3] Institute of Molecular Medicine, The First Sechenov Moscow State Medical University, Moscow, Russian Federation
[4] Moscow Polytechnic University, Moscow, Russian Federation

ABSTRACT

BACKGROUND: The markup of medical image datasets is based on the subjective interpretation of the observed entities by radiologists. There is currently no widely accepted protocol for determining ground truth based on radiologists' reports.

AIM: To assess the accuracy of radiologist interpretations and their agreement for the publicly available dataset "CTLungCa-500", as well as the relationship between these parameters and the number of independent readers of CT scans.

MATERIALS AND METHODS: Thirty-four radiologists took part in the dataset markup. The dataset included 536 patients who were at high risk of developing lung cancer. For each scan, six radiologists worked independently to create a report. After that, an arbitrator reviewed the lesions discovered by them. The number of true-positive, false-positive, true-negative, and false-negative findings was calculated for each reader to assess diagnostic accuracy. Further, the inter-observer variability was analyzed using the percentage agreement metric.

RESULTS: An increase in the number of independent readers providing CT scan interpretations leads to accuracy increase associated with a decrease in agreement. The majority of disagreements were associated with the presence of a lung nodule in a specific site of the CT scan.

CONCLUSION: If arbitration is provided, an increase in the number of independent initial readers can improve their combined accuracy. The experience and diagnostic accuracy of individual readers have no bearing on the quality of a crowd-tagging annotation. At four independent readings per CT scan, the optimal balance of markup accuracy and cost was achieved.

Keywords: X-ray computed tomography; datasets as topic; ground truth; observer variation.

To cite this article
Kulberg NS, Reshetnikov RV, Novik VP, Elizarov AB, Gusev MA, Gombolevskiy VA, Vladzymyrskyy AV, Morozov SP. Inter-observer variability between readers of CT images: all for one and one for all. *Digital Diagnostics*. 2021;2(2):105–118. DOI: https://doi.org/10.17816/DD60622

Received: 11.02.2021 **Accepted:** 07.07.2021 **Published:** 13.07.2021

DOI: https://doi.org/10.17816/DD60622

CT图像解释中结论的可变性：一个为所有和所有为一

Nikolas S. Kulberg[1, 2], Roman V. Reshetnikov[1, 3], Vladimir P. Novik[1], Alexey B. Elizarov[1], Maxim A. Gusev [1, 4], Victor A. Gombolevskiy[1], Anton V. Vladzymyrskyy[1], Sergey P. Morozov[1]

[1] Moscow Center for Diagnostics and Telemedicine, Moscow, Russian Federation
[2] Federal Research Center "Computer Science and Control" of Russian Academy of Sciences, Moscow, Russian Federation
[3] Institute of Molecular Medicine, The First Sechenov Moscow State Medical University, Moscow, Russian Federation
[4] Moscow Polytechnic University, Moscow, Russian Federation

结构简评

理由： 医学图像集的标记在很大程度上依赖于观察到的可疑结构的主观解释。目前，没有推荐的协议用于根据医学描述确定参考数据（ground truth）。

目标： 评估参与编制公开数据集»CTLungCa-500»的放射科医生评估的正确性和一致性，以及确定这些指标与对CT研究进行独立解释的专家数量的关系。

方法： 该数据集包括有患肺癌风险的患者的536项CT研究，其中34名放射科医生参加了该研究。每项CT研究都由六位专家独立解释，之后他们发现的可疑结构由另一位专家进行仲裁。对于每位专家计算真阳性，假阳性，真阴性和假阴性结果的数量，在此基础上评估放射科医生的诊断准确性。为了分析放射科医生的结论之间的一致性，使用了百分比度量。

结果：对CT研究进行独立解释的专家数量的增加在一致性降低的情况下导致其评估的正确性增加。在影响成对研究人员之间结论一致性的因素中，关于CT图像的特定部分中存在肺焦点的观点不一致。

结论：独立的初级解释数量的增加使它们的组合正确性会升高，但需要仲裁，放射科医生的资格对分析的质量没有决定性的价值。从结合解释的正确性及其成本的角度来看，由四名放射科医生进行主要标记是最佳的。

关键词： 计算机断层扫描，数据集，参考数据，结论之间的一致性.

引用本文：
Kulberg NS, Reshetnikov RV, Novik VP, Elizarov AB, Gusev MA, Gombolevskiy VA, Vladzymyrskyy AV, Morozov SP. CT图像解释中结论的可变性：一个为所有和所有为一. *Digital Diagnostics*. 2021;2(2):105–118. DOI: https://doi.org/10.17816/DD60622

收到： 11.02.2021 **接受：** 07.07.2021 **发布日期：** 13.07.2021

论证

2017年S.P. Morozov 和合著者 准备了一个公开可用的数据集“肺部计算机断层扫描的标记结果”，后来称为“CTLung500-Ca”[1, 2]。该套件包含 536 张通过计算机断层扫描 (CT) 检查方法获得的有患肺癌风险的患者胸部的 X 射线图像。 每项研究由六名放射技师独立解释，随后由另一位专家对研究结果进行审查。标记使用了一种对发现进行弱注释的方法，即在CT图像上显示有限数量的病灶，通过指定最大直径封闭球体的坐标进行定位，然后进行聚类[2, 3]。S.P. Morozov 和合著者开发了这样的标记和注释协议，因为放射线技师的解释往往是主观的，并且不能避免错误。 假阳性和假阴性发现的价值同样高的条件下，初级解释仲裁可以提高结论的正确性[4]。请注意，这种仲裁仅在放射线技师犯各种错误时才有效。 根据P.G. Herman和S.J. Hessel两名或两名以上的放射科医生做同样的假阳性发现的可能性很小。然而，有相当一部分的错误通常是由两个或两个以上的专家犯的[5]。因此，独立解释 CT 扫描的放射科医生的数量会显着影响标记和注释的准确性。

目的: 该研究的主要目标是调查 CTLungCa-500 CT 扫描数据库中实际独立解释与所犯错误类型之间的关系，并找到可提升最佳标记精度的 CT 扫描解释协议。 这项研究的第二个目的是研究参与数据集开发的放射科医生的结论是否一致。

研究方法

研究设计

这项工作中，我们分析了一项回顾性多中心观察性研究的数据，该研究致力于研究计算机视觉技术在莫斯科医疗保健系统中的使用前景。

遵从准则

纳入标准: 莫斯科综合诊所的患者，年龄在 50 至 75 岁之间，在医生的推荐下接受了诊断性 CT 扫描，怀疑患有肺癌。

执行条件

根据纳入标准，从统一放射信息服务处下载了3897张CT检查。从这个数字中随机选择了 550 个 CT 检查来创建一个数据集“肺部计算机断层扫描的标记结果”。 14 次 CT 扫描被排除在样本之外，因为它们不符合纳入标准或医疗干预方案。

研究持续时间

该数据集包括2015年1月1日至2017年12月31日进行的CT检查结果。

医疗干预说明

成人患者的推荐扫描参数（身高 170 厘米，体重 70 公斤）：在 120 kV 电压、FOV 350 毫米、切片厚度 ≤ 1.5 毫米、相邻切片之间的距离 ≤ 切片厚度下自动调节管上的电流。 扫描时患者仰卧，扫描方向为横膈膜至肺尖，吸气屏气。 重建内核是特定于特定扫描仪制造商的：对于东芝机器 - FC50、FC51、FC52、FC53、FC07 用于肺，FC07、FC08、FC09、FC17、FC18 用于软组织； 适用于西门子设备 - B70、B75 和 B80； 对于飞利浦设备 - Y-Sharp 和 LONG 用于肺部，SOFT 用于软组织； 用于 GE（通用电气）设备 - 用于肺的 LUN 和用于软组织的 SOFT。

本研究的主要结果

两组志愿放射线技师参与了对研究的标记和注释。第一组的代表（初级专家）由 15 名具有 2 至 10 年以上经验的专家组成，对 CT 扫描进行了初级解释。 根据开发的方法，医生在 CT 图像上搜索大小为 4 到 30 毫米的肺病灶，保留关于肺病灶定位（图像中二维发现中心位置和切片的数量）； 发现的直径； 肺病灶类型（实性、半实性或磨砂玻璃型病灶）。建议医生不要在肺部标记钙化和周围病变，在单次 CT 扫描中不要标记超过五个最大的肺部病变。 为了减少遗漏潜在肺部病变的可能性，每项研究均由六名放射技师独立审查。 然后第二组的一名参与者（仲裁员），由三名10年以上经验的X射线专家组成，检查了第一组X射线专家的标记，评估每个标记的有效性。仲裁员还对发现的病灶进行了恶性评估，将其归类为“恶性”或“良性”根据弗莱什纳协会的建议[6]。

伦理审查

这项研究的数据被用来进行分析，俄罗斯放射科医生和放射科医生协会莫斯科地区分会独

立伦理委员会批准了该协议（2020年2月20日第2-1-II-2020号议定书）。患者在研究过程中执行的所有程序都符合地区和国家研究委员会的标准，以及世界医学会的赫尔辛基宣言和台北宣言。

统计分析

为了确定个别专家的特异性和敏感性，对每位执行初始解释的放射科医师计算真阳性、假阳性、真阴性和假阴性结果的数量。 真正阳性(PI) 被认为是放射科医生和仲裁员对特定区域肺病灶的存在和类型（实性、半实性或磨砂玻璃型压实）的意见一致的情况。 假阳性(False-positive) 案例是仲裁者认为主要专家对给定区域中肺部病灶的存在或类型的评估是错误的。真阴性 (IR) 情况被认可，其中放射科医生没有注意到肿块，根据仲裁者的意见，其他五位主要专家中的一位或多位误认为是局灶性病变。 最后，在仲裁者看来，假阴性 (FN) 病例是放射科医生没有识别出其他五名参与者中的一个或多个正确识别的肺病灶。 分析数据时，我们从仲裁员的判断总是正确的假设出发。敏感性(sensitivity, Se)按公式计算

$$Se = \frac{TP}{(TP + FN)}. \quad (1)$$

特性(specificity, Sp)计算为

$$Sp = \frac{TN}{TN + FP}. \quad (2)$$

为每位参与者确定约登指数 (J):

$$J = Se + Sp - 1. \quad (3)$$

为了计算各种初级专家样本的正确性指标（accuracy, Acc），一些案例被认为是真正阳性的结果，根据仲裁员的意见，至少有一名专家从样本中正确识别了CT图像特定区域的肺灶。 真阴性结果包括样本中至少一名专家没有注意到压实，根据仲裁者的意见，他被研究中的任何其他参与者误认为是肺病灶。 正确性计算为

$$Acc = \frac{(TP + TN)}{(P + O)} \times 100, \quad (4)$$

其中 P - 正确发现的数量，O - 错误发现的数量。

有许多度量来评估一个或多个研究人员的一致性。O. Gerke和合著者使一致性研究系统化的建议中提议使用布兰达-阿尔特曼纳分析[7]。科恩[8]和弗莱斯的卡帕[9]是其他常见的指标。然而由于这些方法的所有优点，它们很难解释，因此这项工作的作者确定了一个最简单的选择。这是研究人员之间的一致性百分比指标，它不考虑随机巧合实验的因素。百分比计算为专家意见（存在、类型）一致的病灶占联合标记病灶总数的比例：

$$Consistency = \frac{Matches}{Matches + Mismatches} \times 100. \quad (5)$$

使用 R 3.6.3 的 dplyr[10]、irr [11]和 ggplot2 [12]包进行统计分析[13]。 准备数据时，我们使用了 Python 3.8.2 中独立开发的脚本[14]。

结果

研究对象

共有 31 名放射科医师参与了 CT 图像的初步解读。研究期间，由于拒绝或无法继续研究，来自 15 名专家的原始队列中的每位放射科医生都被另一名专家取代； 一名参与者被更换两次。 放射线技师的工作量分布不均。 来自原始队列的每位专家平均参与标记和注释 1050 ± 140 个可疑结构。更换它们的放射科医生平均标记了 110 ± 42 个局灶性病变。

根据标注结果，该数据集包括 72 次 CT 扫描，其中放射科医师未发现 4 到 30 毫米的肺病灶，以及 464 次有肺病灶的 CT 扫描，共包含仲裁者确认的 3151 个发现。 其中1761个病变被专家归类为可能是恶性的，445个是良性的，945个封印具有不同的性质（它们包含钙化、脂肪、纤维组织或液体）。

主要研究成果

参与标记的放射科医生的敏感性和特性

处理数据集的过程中，每位放射科医师都被分配了一个三位数的识别号 (ID)。 更换专家的情况下，新参与者将继承他的 ID，并带有一个额外的“+”符号。 敏感性的平均值为 34.9%（95%可信区间 [CI] 30.4-39.4），特异性 - 78.4%（95% CI 74.9-81.9）。这明显低于最低指标，在类似的研究场景中证明D. Ardila和合著者的放

射科医生：分别为 62.5% (95% CI 54.4-70.7) 和 95.3% (95% CI 94.0-96.6)[15]。

这种差异的可能原因是标记条件，根据这些条件，初级专家在图像中标记最多5个点。该建议基于 NELSON 研究的结果，根据该研究结果，原发癌的风险随着病变数量增加到 4 个而增加，但对于具有 5 个或更多病变的患者则降低[16]。 多发性病灶（>5）的情况下，这种方法可能会人为地低估初级专家的诊断准确性，因为它带来了额外的自由度，与每个放射科医生标记的一组特定的病灶有关。这种不确定性可以通过引入另一种分类发现来纠正识别为真阳性病例，其中主要专家在 CT 扫描上标记了至少一个确认的病变。这种评估方案中，初级专家的平均灵敏度为66.2%（95%DI 62.1-69.9），特异性为78.5%（95%DI 72.3-84.8）。标记的目的是创建一个旨在训练人工智能算法的数据集。因此，CT 扫描上的每个可疑结构都值得关注。因此，本论文使用了“方法”一节中的标准来评估诊断准确性。根据这些标准按照约登指数ID 012+ (J = 0.472) 的放射科医生表现出最高的效率，最低 (J = -0.188) - ID 008+ 的专家（表 1）。

研究人数对解释正确性的影响

两位主要专家的解释。此分析中，考虑了 97 次 CT 检查的样本，其中 ID 为 012+ 的放射科医生参与了解释。他所有参与者中表现出最高的约登指数（表 1）。 使用此样本量，获得的所有估计值可能与完整数据集的平均值相差不超过 10%[17]。专家标记的样本包含 53 个实性肺部病变、6 个半实性和 5 个磨砂玻璃封印。此外33个放射学家发现的可疑结构在仲裁中没有得到证实。012+的正确率为65.98%：他正确识别了28个坚实的病灶，避免了33个假阳性病例中的32个，在同一项研究中，其他专家错误地识别了两个坚实的病灶和一个半坚实的病灶，并犯了34个错误。除了他之外，还有一位ID为012的放射科医师，也参与了对样本中全部97次CT检查的评分， 具有最低的约登指数之一（0.058，第 24 位，见表 1）。 这位专家正确识别了32个实性病变、1个半实性、1个磨砂玻璃封印，避免了18个假阳性错误。 研究人员之间的一致性为59.8%，他们估计的综合正确率为 81.44%。不一致的来源是这对夫妇在特定区域（92.3% 的病例）和肺病灶类型（7.7% 的病例）存在可疑结构方面的差异。

CT检查在专科医生中的分布是随机进行的。 因此，研究样本中所有 97 项 CT 研究的解释仅由初级专家 012 和 012+ 进行。 除他们外，还有17名放射技师参加了样本标记（标记病灶数量在括号内为每个）：000（11）、002（54）、003（30）、004（27）、005（18） ），006 (40)，007 (10)，008 (16)，009 (17)，010 (32)，011 (24)，013 (30)，014 (52)，004+ (7)，005+ (10)，011+ (1) 和 014+ (9)，这使得可以将样本中所有研究的第二意见由一位专家表达的情况与人群标签模型进行比较，其中该意见由从一些专家组中随机选择的参与者提供，具有不同的特异性和敏感性。

第一组包括 6 名研究人员（表 2）。 本组平均约登指数为0,078±0.045（最大值0,127，最小值0.001），超过了ID 012参与者的指标（0,058）。尽管如此，估计与012+专家的一致性仅为40.2%，估计的综合正确率为74.23%。这对夫妇的大部分分歧 (97.4%) 的根源在于对肺病灶存在的意见分歧。

重复的类似实验中，我们分析了一组不同的参与者（表3）。第 1 组（表 2）和第 2 组（表 3）的参与者数量和构成不同； 此外，每个人标记的焦点数量分布不均。

第 2 组的平均约登指数为 0.099 ± 0.055（最大值 0.173，最小值 - 0.01）并且高于参与者 012 和第 1 组。两位专家对CT检查的三种解释中，第2组参与者和放射医师012+的评估的一致性和综合正确性也是最高的，分别为71.1%和83.50%。89.3%的病例中，研究者之间的分歧与该区域肺病灶的存在有关，10.7%与肺病灶的类型有关。两位专家在任何组合中的一次评分的平均正确率为79.72±4.87%。

由三个或三个以上的研究人员解释。当分析三个或更多研究者的解释时，所有组包括012和012+研究者三位放射科医生的初步标记和注释，其估计的一致性在32.0-42.3%之间，平均综合准确率为89.18±5.10%。四位独立专家的评估一致性下降到16.5±5.7%，平均综合正确率上升到93.82±3.57%。对于5名放射技师，估计的一致性继续下降到9.8±8.1%，准确率上升到97.94±0.14%。最后我们的实验条件下，六位专家的综合正确率为 100%，一致性为 3.1%（图 1

表 1研究参与者的诊断准确性

ID专家	个别病区指标			
	Se，%	Sp，%	尤登指数	标记焦点数量*
000	39,52	73,17	0,127	1079
001	32,63	79,04	0,117	1068
002	28,25	80,19	0,084	1045
003	44,05	67,75	0,118	1094
004	31,37	68,75	0,001	844
005	33,08	72,76	0,058	1222
006	36,91	71,32	0,082	1085
007	37,31	73,43	0,107	884
008	42,01	68,00	0,100	1227
009	36,79	79,50	0,163	1265
010	38,62	71,16	0,098	1166
011	26,05	79,51	0,056	853
012	33,97	71,88	0,058	1045
013	38,52	77,40	0,159	1028
014	37,16	82,32	0,195	850
000+	31,63	79,17	0,108	194
001+	52,94	82,46	0,354	108
002+	62,50	57,14	0,196	46
003+	60,71	86,21	0,469	86
004+	27,78	86,49	0,143	110
005+	41,49	75,86	0,173	152
006+	31,34	74,14	0,055	125
007+	29,73	85,71	0,154	86
008+	18,99	62,16	-0,188	176
009+	25,76	85,11	0,109	113
010+	25,00	75,36	0,004	145
011+	31,58	93,33	0,249	68
012+	53,85	93,33	0,472	97
013+	34,29	85,71	0,170	77
014+	17,95	100,0	0,179	63
000++	0,00	94,87	-0,051	48

注意：*所有在CT检查中发现的病灶都被考虑在内，并在专家参与的标记中，无论他是否识别它们。

）。 因此，专家评价的正确性与一致性之间存在显著的逆相关关系：r=-0,78，p <0,05。为了证实结论P.G. Herman和S.J. Hessel[5] 在97项研究样本中，当由六名专家解释时，85.7%的假阳性错误仅由一名专家犯下，11.4% - 两名，2.9% - 三个同时犯下。 所有六位专家都正确识别了样本中 8.1% 的阳性结果； 25.8% 的假阴性错误是由六分之一的专家犯下的，8.1% - 两名，8.1% - 三名，19.3% - 四名，30.6% - 五名（图 2）。

标记费

为了从资源管理的角度评估标记的最佳效果，必须考虑使用额外专家进行解释的成本CT扫

表 2第 1 组标记可疑结构的分布

ID研究员	000	002	003	004	005	006
焦点数量	11	54	9	3	11	9

表 3第 2 组中标记的可疑结构的分布

ID研究员	005+	010	003	004	005	006	008	009
焦点数量	10	10	21	9	7	31	8	1

描。因此，可以将正确性的提高与研究注释费用的增加进行比较。

由于志愿放射科医生参与了数据集的标记，他们的工作没有报酬。因此，建议根据专家花费的时间来计算标记成本。平均而言，初级专家在解释一张 CT 图像上花费了 12 分钟，而裁判花费了 4 分钟。 本研究中，在 97 张 CT 图像的研究样本中消除错误 C 的成本计算为给定数量的初级专家在仲裁员参与下进行标记的平均成本与一个标记的成本之差。没有仲裁者参与的放射科医师，除以消除的错误数 (N_{err})：

$$C = \frac{(n \times 12 \times 97 + n \times 4 \times 97) - 12 \times 97}{N_{err}}, \qquad (6)$$

其中 n -主要专家的数量。

专家 012+ 犯了 33 个误报和漏报错误。 通过吸引更多专家和进行仲裁来修复的错误数量，以及相应的消除误差成本见表 4。根据每一个新的初级专家增加错误消除成本42.5±10.7分钟的规律，不包括一分。四位初级专家对数据集进行了标记，随后进行了仲裁，导致所选错误的数量急剧增加，从而降低了成本（表4）。

其他研究成果

由于本研究的设计，每个检查员只解释一次单独的 CT 扫描，本研究没有评估各个放射科医生之间结论的一致性。 专家对评估一致性的平均值为 60.5±5.3%，最小值为 53.1%，最大值为73.0%。

评估初级专家一致性的另一种方法是分析每个放射科医生的阳性发现（图3）。对于每一个原始群体的代表，所确定的热点的最大比例（37.6±5.4%）对应于其他专家无法识别的独特发现（图3，*a*）。然后，按降序排列，调查结果为一 (21.4 ± 2.8%)、二 (14.0 ± 2.0%)、四 (9.5 ± 2.3%)、三 (9 , 2 ± 1.8%) 和五 (8.1 ± 3.1) %) 初级专家。 只有来自原始组（ID 002、004、007和010）的四名X射线学家的一致批准发现率超过10%。请注意，根据本工作中提出的方法计算的尤登指数，这些专家均未包括在领导小组中； 此外，专家 004 是该指标队列中最差的（表 1）。 同时，队列中最大的 Youden 指数 (0.195)的 专家014在阳性结果的一致性方面在他的同事中并不突出（图 3，*a*）。来取代原来的初级专家队伍放射技师中有发现的一致性的不同分布（图 3，*b*）。 确定病灶的最大比例（28.9±18.2%）指示独特的发现。 随后是由两名 (23.3 ± 11.0%)、三名 (13.3 ± 10.7%)、五名 (13.2 ± 11.9%)、六名 (11.5 ± 9, 8%) 和四名 (9.7 ± 7.6%) 专家同时鉴定的发现。这个队列中已经有 8 名放射技师（ID 000+、004+、006+、010+、011+、012+

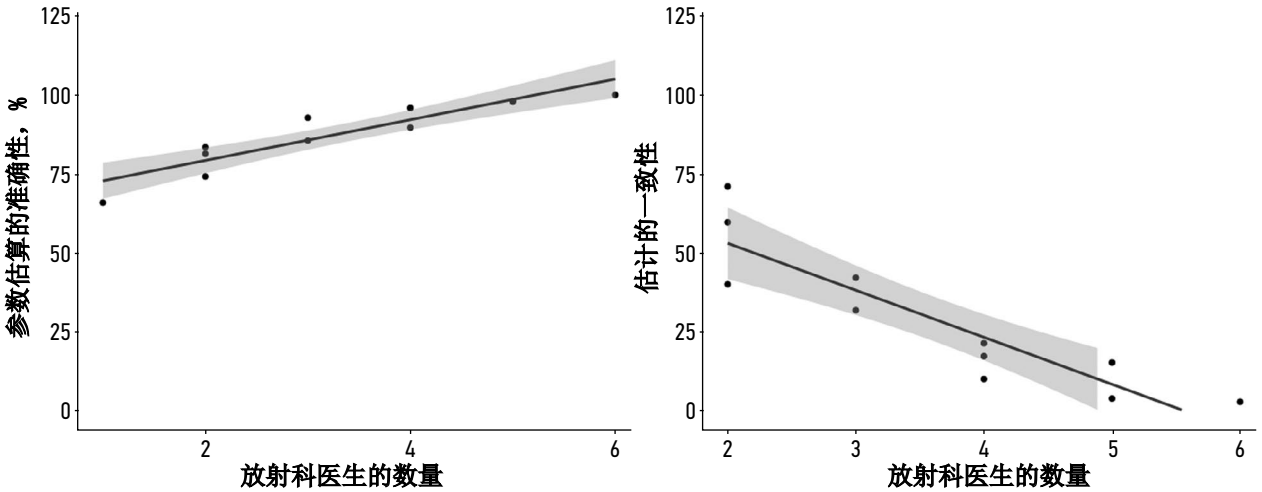

图 1评估的正确性和一致性是参与初级标记的放射科医生数量的函数。灰色表示95%的置信区间。这些点对应于初级专家的不同样本。对于两名、三名和四名专家的实验，从最初的六名放射技师中选择了三个不同的样本； 五 - 两个。

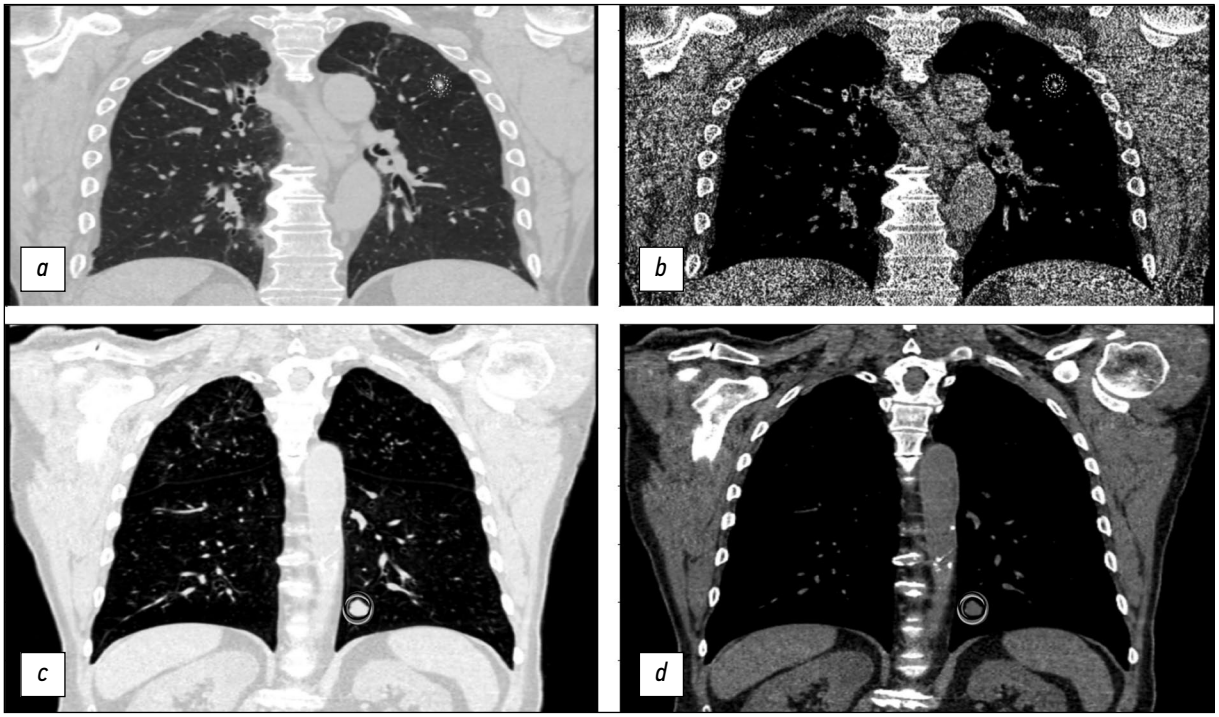

图 2专家之间存在重大分歧的 CT 研究示例（*a*、*b*、CTLungCa-500 AN RLADD02000018919、ID RLSDD02000018855）和完全一致（*c*、*d*、CTLungCa-500 AN RLAD42D007-25151、ID RLSD42500）。研究显示在肺（*a*, *c*）和软组织（*b*, *d*）模式下的正面投影中。垂直分割为 50 毫米，水平分割为 100 像素。放射科医师的标记以不同的颜色显示：*a*、*b* – 六位主要专家中有五位对焦点进行了标记，四位将其指定为实心类型，一位指定为半实心类型。仲裁员不同意他们的意见，认为该发现为良性钙化；*c*, *d* – 所有六位主要评估员和仲裁员都将病变归类为潜在恶性实体。

、013+、014+），其中一致批准的阳性结果比例超过 10%，并且其中四个（ID 000+、010+、011+、014+）超过 20%。 尽管如此这些指标可能是由于该队列中阳性结果的数量很少，这间接证明了他们的一致性的高变异性，以平均值和标准偏差表示。 作为例子是专家 014+，他参与了 CT 研究的解释，其中其他专家确定了 63 个可疑结构（表 1）。 这位专家只标记了七个焦点，其中一个也被另一位专家识别，三个两个，一个五个，两个六个（图 3，*b*）。 同时专家犯了 32 个假阴性错误，从而忽略了约 50% 的真正阳性发现。 于这个队列，在阳性结果的一致性和专家的 尤登分数之间没有观察到相关性。

表 4估计修复错误的成本

初级专家人数	修复的错误数量	成本，最小/错误
2	15	129,3
3	19	183,8
4	29	173,9
5	31	212,8
6	33	246,9

讨论

主要研究成果总结

我们的结果表明，对 CT 检查进行独立解释的专家数量的增加导致他们估计的准确性的提高，并且资格水平不会显着影响放射科医生意见的一致性或他们的综合正确性。影响成对研究人员之间结论一致性的因素中，对于 CT 扫描的特定区域是否存在病变存在意见分歧。

主要研究成果的讨论

目前对于参与医学成像数据集主要标记和注释的放射线技师的推荐数量尚未达成共识。 该值通常在 1 [18, 19]到 4 之间[20]。 我们知道解决这个问题的唯一研究是 P.G.Herman和 S.J.Hessel 认为，随着对研究提供独立解释的专家数量增加，无错误描述的数量逐渐减少[5]。 虽然这当然是一个有趣的观察，但它几乎没有实际价值，因为套利模型原则上基于主要解释将有缺陷的假设。 此外，如果这些错误是不同的，它的效率就会提高。

最后一句话并不总是正确的。 这项工作的结果表明放射科医生犯下不同的错误并不能自动

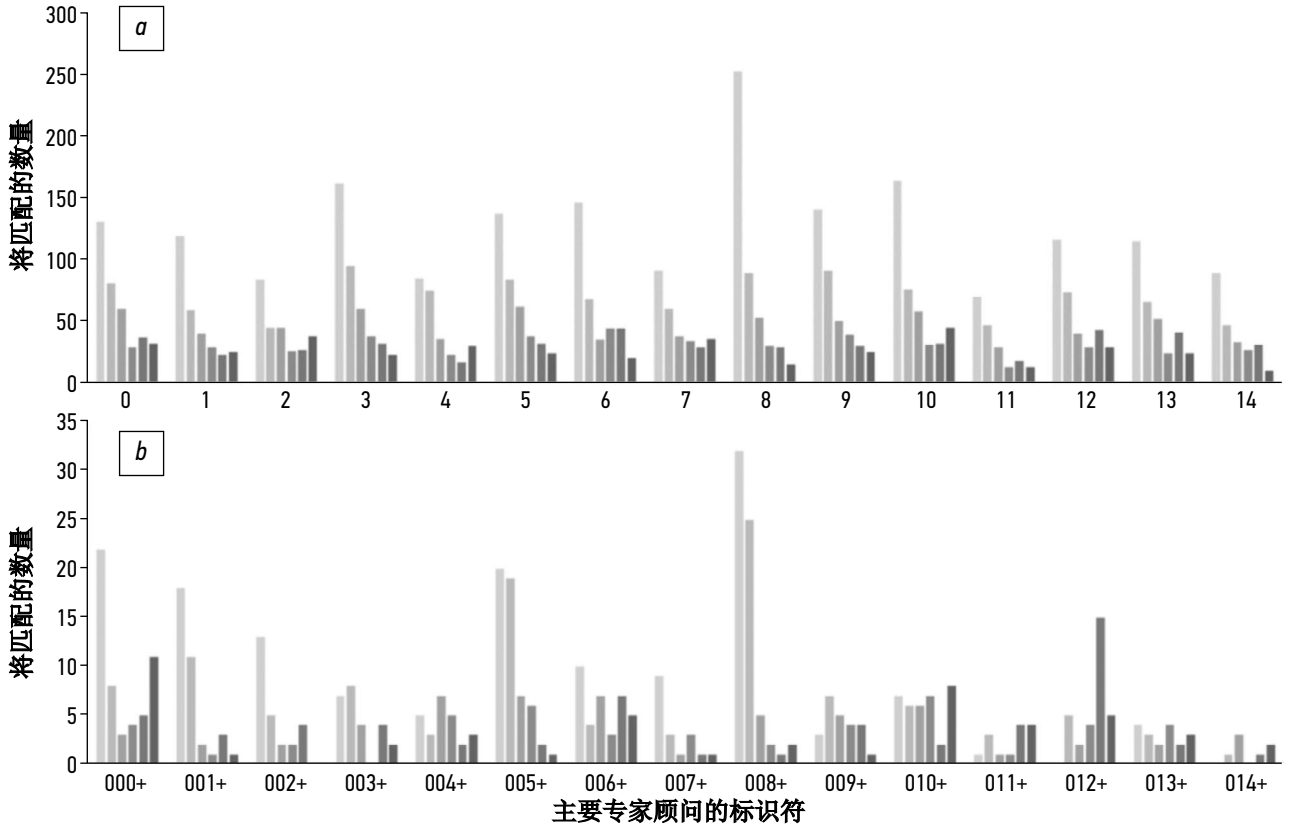

图 3主要专家之间的协议： *a* – 原始 15 名放射线技师的代表； *b* – 更换放射线技师。由于注意到的病变数量很少，因此没有给出 ID 000++ 专家的数据。对于每位放射科医生，第一列对应于该专家唯一标记的病变数量（其他五位专家均未识别出这一发现）。以下列对应于放射科医师确定的病变由一名、两名、三名、四名和五名其他主要专家注意到的情况。该图没有考虑仲裁员的批准，以及放射科医师对病变类型的意见分歧。

提高他们结论的组合正确性。由两位对 CT 图像进行主要解释的专家进行的实验中，在第二对中观察到了最高程度的不一致（一致 40.2%），但它也证明了三对中的正确性最低（74.2 对 81.4 和 83.5 %） 。 同时，第三对显示出最高的准确度值，具有最大的一致性（71.1%）。尽管如此，根据这项工作获得的数据，专家评估的一致性与其正确性之间存在显着的负相关（r = -0.78）。 因此，在两名放射技师的初步解释中，一致率为 57.0 ± 15.6%，正确率为 79.7 ± 4.9%； 对于五位放射技师，这些指标分别等于 9.8 ± 8.1% 和 97.9 ± 0.1%，并且这种依赖性在所有考虑的标记数据集变体中都保留了下来（图 1）。

根据本研究的结果，通过四位初级专家和后续仲裁的方法可以实现正确性和加价成本的最佳组合（表4）。 对他来说，消除错误的次数比三位放射科医师的打分明显增加，同时消除一个错误所用的时间减少（-9.9分钟）。 额外的初级专家的参与导致解释的正确性进一步提高。然而，这是由于消除错误的成本平均增加了 42.5 ± 10.7 分钟。

这项工作中，当将主要专家的评估分配给假阴性、真阴性、假阳性和真阳性的类别时，是基于所有肺病灶都将被标记的假设在每次 CT 扫描中。 然而研究结果表明，研究参与者遵循他们的建议，将自己限制在 CT 扫描中最大的五个肺部病变。因此，个体放射技师忽略了很大一部分肺部病变，这影响了他们的诊断准确率，以及专家对的一致性值。 然而，主要专家的意见分歧是仲裁的可取结果，因为它扩大了所报告的可疑实体的目录。即使在人为限制要标记的病灶数量的情况下，这也减少了假阴性结果的比例。 这项工作的主要发现之一是，多位放射技师之间的共识并不是对数据集进行良好标记的先决条件。主要责任在于仲裁员，他们必须正确解释主要专家指出的所有可疑结构 （图2,a,b）

研究的局限性

这项工作的主要限制是确定参考数据（ground

truth）的模型 – 那些应该被视为肺病灶的发现。 解释 CT 扫描时，放射科医生无法访问患者的临床、生物学和基因组数据； 此外，对于所有患者，该集合不包含两个在时间上间隔开的研究，这将使评估可疑结构发展的动态成为可能。 我们也从仲裁员的意见总是正确的假设出发，我们将主要意见与仲裁员意见之间的分歧解释为总是支持后者。 然而，该试剂盒包含的许多例子让人怀疑这种方法的可靠性：特别是，仲裁者将 19 个肺部病变标记为良性和恶性。S.J. Hessel和合著者提议仲裁员只能正确解决主要专家之间约 80% 的分歧[4]。

这项工作的另一个限制是无法评估个别放射线技师结论的可重复性。为了实现研究的主要目标，采用了有限样本；为了更可靠的统计，最好的方法是使用样本复制方法(bootstrap)。最后，本研究中主要检查员诊断准确性的评估依赖于他们将标记所有肺病灶的假设。 如果 CT 扫描的病灶数量超过 5 个，则该假设与标记建议相冲突，这可能会影响敏感性和特异性的最终个体指标。为了弥补这种方法学上的局限性，研究作者试图评估由两名、三名、四名和五名其他放射线技师批准的每位初级检查员的阳性结果数量的一致性（图 3）。 这样的分析没有考虑假阴性错误，因此其结果与每个专家的约登指数获得的值不相关。 最重要的是，这项研究检查了全剂量 CT 扫描的解释结果。因此，她的结论可能不适用于筛查研究中获得的数据，这些研究的特点是使用低剂量和超低剂量 CT 协议。

结论

尽管有其局限性，这项工作令人信服地表明，增加独立的主要解释的数量可以增加其正确性，但须经仲裁。同时，放射技师的资格对分析的质量并不是决定性的，因为根据获得的结果，他们的评估的综合正确性并不取决于个别尤登指数。由四位专家对CT检查进行初步独立解释的过程中，实现了正确性和标记成本的最佳组合。这一观察结果为开发人工智能算法的需求创造了理论基础，这些算法旨在通过在CT扫描上标记可疑结构和引导放射科医生的注意力来诊断疾病。此外，这项工作中获得的结果使我们有可能证实数据集多用户标记（crowd-tagging）的项目模型，在这种模型中，标记数量的增加将导致一致性的降低，并同时提高仲裁提供的最终产品的质量。

附加信息

资金来源。作者声称这项研究没有资金支持。

利益冲突。本文作者已证实没有利益冲突需要报道。

作者贡献。所有作者都确认其作者符合国际ICMJE标准（所有作者为文章的概念，研究和准备工作做出了重大贡献，并在发表前阅读并批准了最终版本）。最大的贡献分布如下： N.S. Cullberg – 数据集设计、研究概念化、文章准备和编辑； R.V. Reshetnikov – 统计分析，撰写文章正文；V.P. Novik – 准备数据集，编写用于收集数据的脚本，统计分析； A.B. Elizarov – 准备数据集，编写脚本来收集数据；M.A. Gusev – 准备数据集，编写脚本来收集数据； V.A. Gombolevsky – 研究概念化、数据集设计；A.V. Vladzimirskiy – 研究的概念化，编辑文章的文本； S.P. Morozov – 数据集设计、概念化和研究资金。

谢意的表示。作者对切尔尼娜•瓦莱里娅•尤里耶夫娜方法学咨询以及所有参与数据集标记的放射科医生表示最深切的感谢。

СПИСОК ЛИТЕРАТУРЫ

1. Морозов С.П., Кульберг Н.С., Гомболевский В.А., и др. Датасет радиологии Москвы CTLungCa-500. 2018. Режим доступа: https://mosmed.ai/datasets/ct_lungcancer_500/. Дата обращения: 11.02.2021.

2. Morozov S.P., Gombolevskiy V.A., Elizarov A.B., et al. A simplified cluster model and a tool adapted for collaborative labeling of lung cancer CT Scans // Comput Methods Programs Biomed. 2021. Vol. 206. P. 106111. doi: 10.1016/j.cmpb.2021.106111

3. Kulberg N.S., Gusev M.A., Reshetnikov R.V., et al. Methodology and tools for creating training samples for artificial intelligence systems for recognizing lung cancer on CT images // Heal Care Russ Fed. 2020. Vol. 64, N 6. P. 343–350. doi: 10.46563/0044-197X-2020-64-6-343-350

4. Hessel S.J., Herman P.G., Swensson R.G. Improving performance by multiple interpretations of chest radiographs: effectiveness and cost // Radiology. 1978. Vol. 127, N 3. P. 589–594. doi: 10.1148/127.3.589

5. Herman P.G., Hessel S.J. Accuracy and its relationship to experience in the interpretation of chest radiographs // Invest Radiol. 1975. Vol. 10, N 1. P. 62–67. doi: 10.1097/00004424-197501000-00008
6. MacMahon H., Naidich D.P., Goo J.M., et al. Guidelines for management of incidental pulmonary nodules detected on ct images: from the fleischner society 2017 // Radiology. 2017. Vol. 284, N 1. P. 228–243. doi: 10.1148/radiol.2017161659
7. Gerke O., Vilstrup M.H., Segtnan E.A., et al. How to assess intra- and inter-observer agreement with quantitative PET using variance component analysis: a proposal for standardisation // BMC Med Imaging. 2016. Vol. 16, N 1. P. 54. doi: 10.1186/s12880-016-0159-3
8. Rasheed K., Rabinowitz Y.S., Remba D., Remba M.J. Interobserver and intraobserver reliability of a classification scheme for corneal topographic patterns // Br J Ophthalmol. 1998. Vol. 82, N 12. P. 1401–1406. doi: 10.1136/bjo.82.12.1401
9. Van Riel S.J., Sánchez C.I., Bankier A.A., et al. Observer variability for classification of pulmonary nodules on low-dose ct images and its effect on nodule management // Radiology. 2015. Vol. 277, N 3. P. 863–871. doi: 10.1148/radiol.2015142700
10. Wickham H., François R., Henry L., Müller K. dplyr: A Grammar of Data Manipulation. R package version 1.0.4. 2021.
11. Gamer M, Lemon J, Fellows I, Singh P. irr: Various Coefficients of Interrater Reliability and Agreement. 2019.
12. Wickham H. ggplot2: elegant Graphics for Data Analysis. Springer-Verlag New York; 2016. 260 p.
13. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. 2020. Режим доступа: http://www.r-project.org/index.html. Дата обращения: 11.02.2021.
14. Van Rossum G., Drake F.L. Python 3 Reference Manual. CreateSpace, Scotts Valley, CA; 2009.
15. Ardila D., Kiraly A.P., Bharadwaj S., et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography // Nat Med. 2019. Vol. 25, N 6. P. 954–961. doi: 10.1038/s41591-019-0447-x
16. Peters R., Heuvelmans M., Brinkhof S., et al. Prevalence of pulmonary multi-nodularity in CT lung cancer screening. 2015.
17. Creative Research Systems. The survey systems: Sample size calculator. 2012.
18. Hugo G.D., Weiss E., Sleeman W.C., et al. A longitudinal four-dimensional computed tomography and cone beam computed tomography dataset for image-guided radiation therapy research in lung cancer // Med Phys. 2017. Vol. 44, N 2. P. 762–771. doi: 10.1002/mp.12059
19. Bakr S., Gevaert O., Echegaray S., et al. A radiogenomic dataset of non-small cell lung cancer // Sci Data. 2018. Vol. 5. P. 180202. doi: 10.1038/sdata.2018.202
20. Armato S.G., McLennan G., Bidaut L., et al. The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on ct scans // Med Phys. 2011. Vol. 38, N 2. P. 915–931. doi: 10.1118/1.3528204

REFERENCES

1. Morozov SP, Kulberg NS, Gombolevsky VA, et al. Moscow Radiology Dataset CTLungCa-500. 2018. (In Russ). Available from: https://mosmed.ai/datasets/ct_lungcancer_500/
2. Morozov SP, Gombolevskiy VA, Elizarov AB, et al. A simplified cluster model and a tool adapted for collaborative labeling of lung cancer CT Scans. *Comput Methods Programs Biomed*. 2021;206:106111. doi: 10.1016/j.cmpb.2021.106111
3. Kulberg NS, Gusev MA, Reshetnikov RV, et al. Methodology and tools for creating training samples for artificial intelligence systems for recognizing lung cancer on CT images. *Heal Care Russ Fed*. 2020;64(6):343–350. doi: 10.46563/0044-197X-2020-64-6-343-350
4. Hessel SJ, Herman PG, Swensson RG. Improving performance by multiple interpretations of chest radiographs: effectiveness and cost. *Radiology*. 1978;127(3):589–594. doi: 10.1148/127.3.589
5. Herman PG, Hessel SJ. Accuracy and its relationship to experience in the interpretation of chest radiographs. *Invest Radiol*. 1975;10(1):62–67. doi: 10.1097/00004424-197501000-00008
6. MacMahon H, Naidich DP, Goo JM, et al. Guidelines for management of incidental pulmonary nodules detected on ct images: from the fleischner society 2017. *Radiology*. 2017;284:228–243. doi: 10.1148/radiol.2017161659
7. Gerke O, Vilstrup MH, Segtnan EA, et al. How to assess intra- and inter-observer agreement with quantitative PET using variance component analysis: a proposal for standardisation. *BMC Med Imaging*. 2016;16(1):54. doi: 10.1186/s12880-016-0159-3
8. Rasheed K, Rabinowitz YS, Remba D, Remba MJ. Interobserver and intraobserver reliability of a classification scheme for corneal topographic patterns. *Br J Ophthalmol*. 1998;82(12):1401–1406. doi: 10.1136/bjo.82.12.1401
9. Van Riel SJ, Sánchez CI, Bankier AA, et al. Observer variability for classification of pulmonary nodules on low-dose ct images and its effect on nodule management. *Radiology*. 2015;277(3):863–871. doi: 10.1148/radiol.2015142700
10. Wickham H, François R, Henry L, Müller K. dplyr: A Grammar of Data Manipulation. R package version 1.0.4. 2021.
11. Gamer M, Lemon J, Fellows I, Singh P. irr: Various Coefficients of Interrater Reliability and Agreement. 2019.
12. Wickham H. ggplot2: elegant Graphics for Data Analysis. Springer-Verlag New York; 2016. 260 p.
13. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria; 2020. Available from: http://www.r-project.org/index.html
14. Van Rossum G, Drake FL. Python 3 Reference Manual. CreateSpace, Scotts Valley, CA; 2009.
15. Ardila D, Kiraly AP, Bharadwaj S, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat Med*. 2019;25(6):954–961. doi: 10.1038/s41591-019-0447-x
16. Peters R, Heuvelmans M, Brinkhof S, et al. Prevalence of pulmonary multi-nodularity in CT lung cancer screening. 2015.
17. Creative Research Systems. The survey systems: Sample size calculator. 2012.
18. Hugo GD, Weiss E, Sleeman WC, et al. A longitudinal four-dimensional computed tomography and cone beam computed tomography dataset for image-guided radiation therapy research in lung cancer. *Med Phys*. 2017;44(2):762–771. doi: 10.1002/mp.12059
19. Bakr S, Gevaert O, Echegaray S, et al. A radiogenomic dataset of non-small cell lung cancer. *Sci Data*. 2018;5:180202. doi: 10.1038/sdata.2018.202

20. Armato SG, McLennan G, Bidaut L, et al. The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on ct scans. *Med Phys*. 2011;38(2):915–931. doi: 10.1118/1.3528204

ОБ АВТОРАХ

*** Кульберг Николай Сергеевич,** к.ф.-м.н.;
адрес: Россия, 127051, Москва, ул. Петровка, д. 24;
ORCID: https://orcid.org/0000-0001-7046-7157;
eLibrary SPIN: 2135-9543; e-mail: kulberg@npcmr.ru

Решетников Роман Владимирович, к.ф.-м.н.;
ORCID: https://orcid.org/0000-0002-9661-0254;
eLibrary SPIN: 8592-0558; e-mail: reshetnikov@fbb.msu.ru

Новик Владимир Петрович;
ORCID: https://orcid.org/0000-0002-6752-1375;
eLibrary SPIN: 2251-1016; e-mail: v.novik@npcmr.ru

Елизаров Алексей Борисович, к.ф.-м.н.;
ORCID: https://orcid.org/0000-0003-3786-4171;
eLibrary SPIN: 7025-1257; e-mail: a.elizarov@npcmr.ru

Гусев Максим Александрович;
ORCID: https://orcid.org/0000-0001-8864-8722;
eLibrary SPIN: 1526-1140; e-mail: m.gusev@npcmr.ru

Гомболевский Виктор Александрович, к.м.н.;
ORCID: https://orcid.org/0000-0003-1816-1315;
eLibrary SPIN: 6810-3279; e-mail: g_victor@mail.ru

Владзимирский Антон Вячеславович, д.м.н., профессор;
ORCID: https://orcid.org/0000-0002-2990-7736;
eLibrary SPIN: 3602-7120; e-mail: a.vladzimirsky@npcmr.ru

Морозов Сергей Павлович, д.м.н., профессор;
ORCID: https://orcid.org/0000-0001-6545-6170;
eLibrary SPIN: 8542-1720; e-mail: morozov@npcmr.ru

* Автор, ответственный за переписку / Corresponding author

AUTHORS' INFO

*** Nikolas S. Kulberg,** Cand. Sci. (Phys.-Math.);
address: 24 Petrovka str., 109029, Moscow, Russia;
ORCID: https://orcid.org/0000-0001-7046-7157;
eLibrary SPIN: 2135-9543; e-mail: kulberg@npcmr.ru

Roman V. Reshetnikov, Cand. Sci. (Phys.-Math.);
ORCID: https://orcid.org/0000-0002-9661-0254;
eLibrary SPIN: 8592-0558; e-mail: reshetnikov@fbb.msu.ru

Vladimir P. Novik;
ORCID: https://orcid.org/0000-0002-6752-1375;
eLibrary SPIN: 2251-1016; e-mail: v.novik@npcmr.ru

Alexey B. Elizarov, Cand. Sci. (Phys.-Math.);
ORCID: https://orcid.org/0000-0003-3786-4171;
eLibrary SPIN: 7025-1257; e-mail: a.elizarov@npcmr.ru

Maxim A. Gusev;
ORCID: https://orcid.org/0000-0001-8864-8722;
eLibrary SPIN: 1526-1140; e-mail: m.gusev@npcmr.ru

Victor A. Gombolevskiy, MD, Cand. Sci. (Med.);
ORCID: https://orcid.org/0000-0003-1816-1315;
eLibrary SPIN: 6810-3279; e-mail: g_victor@mail.ru

Anton V. Vladzymyrskyy, MD, Dr. Sci. (Med.), Professor;
ORCID: https://orcid.org/0000-0002-2990-7736;
eLibrary SPIN: 3602-7120; e-mail: a.vladzimirsky@npcmr.ru

Sergey P. Morozov, MD, Dr. Sci. (Med.), Professor;
ORCID: https://orcid.org/0000-0001-6545-6170;
eLibrary SPIN: 8542-1720; e-mail: morozov@npcmr.ru