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

ЭКО•ВЕКТОР

Лицензия CC BY-NC-ND 4.0
© Коллектив авторов, 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

ECO•VECTOR

The article can be use under the CC BY-NC-ND 4.0 license
© Authors, 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

ECO•VECTOR
The article can be use under the CC BY-NC-ND 4.0 license
© Authors, 2021

INTRODUCTION

In 2017, S.P. Morozov et al. prepared a publicly available dataset, "Tagged results of computed tomography of the lungs," later called "CTLung500-Ca" [1, 2]. This set comprises 536 computed tomography (CT) chest X-ray images of lung cancer high risk patients. Each study was independently interpreted by six radiographers, and the findings were subsequently reviewed by an additional expert. The markup used an approach with a weak annotation of findings, i.e., the indication of a limited number of nodules on the CT image, which were localized by specifying the coordinates of the enclosing spheres of maximum diameter with their subsequent clustering [2, 3]. S.P. Morozov et al. developed such a markup and annotation protocol because the interpretations of radiologists tend to be subjective and are not immune to error. Under conditions in which the costs of false positive (FP) and false negative (FN) findings are equally high, the arbitration of primary interpretations can increase the correctness of conclusions [4]. Such arbitration is only effective if radiographers commit different mistakes. According to P.G. Herman and S.J. Hessel, the probability that two or more radiographers can make the same FP finding is low. However, a significant proportion of FN errors, as a rule, is made by two or more specialists [5]. Thus, the number of radiologists who independently interpret CT scans can affect significantly the correctness of markup and annotation.

STUDY AIM

The study primarily aimed to investigate the relationship between the number of independent interpretations located in the CTLungCa-500 CT scan database and the number and type of errors made and to search for a CT scan interpretation protocol that promotes optimal tagging correctness. The secondary aim of the study was the analysis of agreement between the radiographers who participated in the dataset preparation.

METHODS

Study design

In this work, we analyzed the data of a retrospective multicenter observational study focused on the analysis of prospects for the use of computer vision technologies in the healthcare system of Moscow.

Inclusion criteria

The inclusion criteria were patients of polyclinics in Moscow, aged 50–75 years, who underwent a diagnostic CT study referred by an attending physician due to suspected lung cancer.

Conditions in conducting the experiment

In accordance with the inclusion criteria, 3897 CT examinations were downloaded from the Unified Radiological Information Service. A total of 550 CT examinations were selected randomly from this array to create a dataset, "Tagged results of computed tomography of the lungs." Exactly 14 CT scans were excluded from the sample due to non-compliance with the inclusion criteria or the protocol of medical intervention.

Study duration

The dataset included the results of CT examinations conducted from January 01, 2015 to December 31, 2017.

Description of the medical intervention

The recommended scanning parameters for adult patients (height: 170 cm, body weight: 70 kg) included the automatic modulation of the current on the tube at a voltage of 120 kV, field of view of 350 mm, slice thickness of 1.5 mm or less, and the distance between adjacent slices the same as the slice thickness or less. Scanning was performed with the patient in the supine position, with the scanning directed from the diaphragm to the apex of the lungs within a single breath-hold. Reconstruction kernels were specific for a particular tomographic scanner manufacturer, namely, FC50, FC51, FC52, FC53, and FC07 for lungs and FC07, FC08, FC09, FC17, and FC18 for soft tissues for Toshiba machines; B70, B75, and B80 for Siemens devices; Y-Sharp and LUNG for lungs and SOFT for soft tissues for Philips devices; LUNG for lungs and SOFT for soft tissues for GE (General Electrics) devices.

Primary study outcome

Two groups of volunteer radiographers participated in the tagging and annotation of the studies. Representatives of Group 1 (primary experts), consisting of 15 specialists with working experience of 2–10 years or more, performed the primary interpretation of CT scans. In accordance with the developed methodology, doctors searched for pulmonary nodules with sizes from 4 mm to 30 mm on CT images and retained the information about the findings, such as localization of pulmonary nodules (position of the center of the finding by defined by two dimensions in the image and the slice number); diameter of the finding; type of pulmonary nodule (solid, part solid, or ground glass opacity nodule). Medical specialists were advised not to mark calcified and peri-fissural lesions in the lungs and not to mark more than five of the largest pulmonary nodules on a single CT scan. Each study was reviewed independently by six radiographers to reduce the probability of missing potential pulmonary lesions. Then, one of the participants in Group 2 (arbitrators), consisting of three radiologists with 10 or more years of working experience, reviewed the tagging made by the radiologists of Group 1 to assess the significance of each mark. The arbitrators also assessed the malignancy of the lesions detected, referring them to the category of "malignant" or "benign," guided by the Fleischner Society recommendations [6].

Ethical considerations

The study, whose data were used for the analysis in this work, was approved by the Independent Ethics Committee of the Moscow Regional Branch of the Russian Society of Roentgenologists and Radiologists (Protocol No. 2 1-II-2020 dated February 20, 2020). All procedures performed on patients during the study were in accordance with the standards of the regional and national research committee and the Declaration of Helsinki and the Taipei Declaration of the World Medical Association.

Statistical analysis

The numbers of true positive (TP), FP, true negative (TN), and FN findings were counted for each radiologist who performed the initial interpretation to determine the specificity (Sp) and sensitivity (Se) of individual specialists. The cases were considered TP if the opinions of the radiologist and the arbitrator coincided about the presence and type of a pulmonary nodule (solid, part solid, or ground glass) in a particular area. The cases were FP if the arbitrator recognized the primary expert's assessment as erroneous regarding the presence or type of a pulmonary nodule in a given area. The cases were considered TN when the radiologist did not mark the entity, which in the opinion of the arbitrator, was mistaken for a lung nodule by one or more of the other five primary experts. Finally, for FN cases, the radiologist did not recognize a pulmonary nodule that was correctly identified by one or more of the five other participants, in the opinion of the arbitrator. When analyzing the data, we assumed that the arbitrator's opinion is always correct.

Se was calculated by the following equation:

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

Sp was calculated as follows:

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

For each participant, Youden's index (J) was determined:

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

To calculate the accuracy indicator (*Acc*) of different samples of primary experts, we defined the TP as the cases when at least one specialist from the sample identified correctly, in the opinion of the arbitrator, a pulmonary nodule in a specific area of the CT scan. The TN results included cases in which at least one specialist from the sample did not notice a lesion, which was mistaken, in the opinion of the arbitrator, for a pulmonary nodule by any other participant in the study. The accuracy was calculated as follows

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

where P is the number of correct findings, and N is the number of incorrect findings.

A number of metrics are available for the assessment of agreement among one or more researchers. O. Gerke et al., in their recommendations for the systematization of agreement studies, suggested using the Bland–Altman analysis [7]. Other common metrics are Cohen's [8] and Fleiss' [9] kappa. However, with all the advantages of these methods, they are difficult to interpret. Thus, the authors of this work settled on the simplest option, that is, the percentage agreement between researchers, which disregards the factor of random coincidences of radiologists' conclusions but at the same time is intuitively comprehensible and reflects reliably the main regularities, provided that repeated experiments are performed. The percentage was calculated as the proportion of nodules for which expert opinions (presence, type) coincided in relation to the total number of jointly tagged nodules:

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

Statistical analysis was performed using the dplyr [10], irr [11], and ggplot2 [12] packages for R 3.6.3 [13]. When preparing the data, we used self-written scripts in the Python 3.8.2 language [14].

RESULTS

Research objects

A total of 31 radiologists took part in the primary interpretation of CT images. Each radiologist from the initial cohort of 15 specialists was replaced by another specialist during the study due to refusal or inability to continue the study; one participant was replaced twice. The radiographers' workload was distributed unevenly. Each specialist from the initial cohort participated in labeling and annotating an average of 1050 ± 140 lesions. The radiologists who replaced them tagged an average of 110 ± 42 lesions.

Based on the tagging results, the dataset included 72 CT scans, in which radiologists did not find pulmonary nodules from 4 mm to 30 mm, and 464 CT scans with pulmonary nodules, comprising 3151 findings confirmed by the arbitrator. A total of 1761 lesions were classified by experts as presumable malignant, 445 lesions as benign, and 945 entities of a different nature (they contained calcifications, adipose tissue, fibrous tissue, or fluid).

Key research findings

Se and Sp of radiographers involved in the tagging

During the work on the dataset, a three-digit identification number (ID) was assigned to each radiologist. In the case of replacement of a specialist, the new participant inherited his ID with an additional "+" symbol. The average value of Se was 34.9% (95% confidence interval [CI]: 30.4–39.4), and that of Sp was 78.4% (95% CI: 74.9–81.9),

which was noticeably inferior to the minimum indicators demonstrated by radiologists in a similar study of D. Ardila et al., namely, 62.5% (95% CI: 54.4–70.7) and 95.3% (95% CI: 94.0–96.6), respectively [15].

The difference noted was possibly caused by the tagging recommendations, guided by which the primary experts tagged a maximum of five nodules in the image. This recommendation is based on the results of the NELSON study, according to which the risk of primary cancer increases with increase in the number of lesions to four but decreases for patients with five or more lesions [16]. In cases of multiple lesions (>5), this approach can artificially underestimate the diagnostic accuracy of primary experts because it introduces an additional degree of freedom associated with a specific set of lesions that each radiologist has tagged. This uncertainty can be corrected by introducing an alternative classification of findings, recognizing the cases as TP when the primary expert tagged at least one confirmed nodule on the CT scan. With this assessment scheme, the average Se of primary experts was 66.2% (95% CI: 62.1–69.9), and the Sp was 78.5% (95% CI: 72.3–84.8). However, the markup was aimed at creating a dataset designed to train artificial intelligence algorithms, and every suspicious structure on a CT image was of interest. For this reason, in this work, the criteria set out in the Methods section were used to assess the diagnostic accuracy. In accordance with these criteria and based on Youden's index, the radiologist with ID 012+ showed the highest accuracy (J = 0.472), and the specialist with ID 008+ had the owest (J = –0.188) (Table 1).

Table 1. Diagnostic correctness of study participants.

Expert ID	Indicators for individual nodules			
	Se, %	Sp, %	Youden's Index	Number of tagged nodules*
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

Note. *All lesions revealed in CT examinations were considered in the tagging in which the expert participated, regardless of whether he recognized them or not.

Influence of the number of researchers on the interpretation accuracy

Interpretation by two primary experts. In this analysis, a sample of 97 CT studies was considered and interpreted by the radiologist (ID 012+) who showed the highest Youden's index score among all participants (Table 1). With this sample size, all estimates obtained may differ from the average for the full data set by no more than 10% [17]. The sample tagged by this specialist contained 53 solid pulmonary lesions, 6 part solid, and 5 ground glass lesions. In addition, 33 entities discovered by radiologists were not confirmed in the course of arbitration. The accuracy of assessments by Radiologist 012+ was 65.98%, that is, he correctly identified 28 solid nodules and avoided 32 out of 33 FP errors made by other specialists in the same studies while recognizing incorrectly 2 solid and 1 part solid nodules and committing 34 FN errors. In addition, the radiologist with ID 012, who had one of the lowest Youden's index scores (0.058, place 24; Table 1), also participated in tagging all 97 CT studies in the sample. This specialist correctly recognized 32 solid lesions, 1 part solid, and 1 ground glass lesion and avoided 18 FP errors. With the agreement between researchers equaling 59.8%, the joint accuracy of their estimates was 81.44%. The sources of disagreement were the discrepancy between the opinions within the pair regarding the presence of a lesion in a particular area (92.3% of cases) and the type of pulmonary nodule (7.7% of cases).

The distribution of CT studies among specialists was performed in a random manner. For this reason, all 97 CT studies in the studied sample were interpreted only by primary Experts 012 and 012+. In addition, 17 radiographers participated in sample tagging (the number of tagged nodules is indicated in the brackets for each ID), namely, 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), and 014+(9). They enabled the comparison of the situation in which the second opinion on all studies in the sample was expressed by one specialist, with the crowd-tagging model, in which an opinion is provided by a participant selected randomly from a certain expert group with variable Sp and Se indices.

Group 1 included six researchers (Table 2). The average Youden's index in this group was 0.078 ± 0.045 (maximum value: 0.127; minimum value: 0.001), which exceeded the indicator of Radiologist ID 012 (0.058). Nevertheless, the agreement of estimates with Radiologist 012+ was 40.2%, and the joint accuracy of the estimates was 74.23%. The source of most of disagreements in the pair (97.4%) was the divergence of opinions about the presence of pulmonary nodules.

In a repeated similar experiment, a group with a different composition of participants was analyzed (Table 3). The number and composition of participants differed between Groups 1 (Table 2) and 2 (Table 3). Moreover, the distribution of the number of nodules tagged by each expert was uneven.

The mean Youden's index in Group 2 was 0.099 ± 0.055 (maximum: 0.173, minimum: 0.01) and was higher than that by Radiologist 012 and in Group 1. The agreement and joint accuracy of the assessments of participants in Group 2 and Radiologist 012+ were the highest of the three considered options for the interpretation of CT studies by two experts, accounting for 71.1% and 83.50%, respectively. The disagreement between researchers in 89.3% of cases was associated with the presence of a pulmonary nodule in this area and with its type in 10.7%. The average accuracy of interpretations during the primary tagging by two specialists in any combination was 79.72% ± 4.87%.

Interpretation by three or more researchers. When analyzing the interpretation by three or more researchers, all groups included Radiologists 012 and 012+. With the primary tagging and annotation by three radiologists, the agreement of their estimates ranged from 32.0% to 42.3%, and the average joint accuracy was 89.18% ± 5.10%. The inter-observer agreement between the assessments of four independent specialists decreased to 16.5% ± 5.7%, whereas the average joint accuracy increased to 93.82% ± 3.57%. For five radiographers, the inter-observer agreement continually declined to 9.8% ± 8.1%, and the accuracy continually increased to 97.94% ± 0.14%. Finally, the joint accuracy of the six experts was 100% under our experimental conditions, with the agreement of 3.1% (Fig. 1). Thus, a significant inverse correlation existed between the accuracy and agreement of expert assessments ($r = -0.78$, $p < 0.05$).

In support of the conclusions by P.G. Herman and S.J. Hessel [5], in a sample of 97 studies, when interpreted by six specialists, 85.7% of FP errors were made by one expert, 11.4% by two experts, and 2.9% by three experts at the same time. All six experts identified correctly 8.1% of positive findings in the sample. Meanwhile, 25.8% of FN errors

Table 2. Distribution of tagged suspicious structures in Group 1.

Researcher ID	000	002	003	004	005	006
Number of tagged nodules	11	54	9	3	11	9

Table 3. Distribution of tagged suspicious structures in Group 2.

Researcher ID	005+	010	003	004	005	006	008	009
Number of tagged nodules	10	10	21	9	7	31	8	1

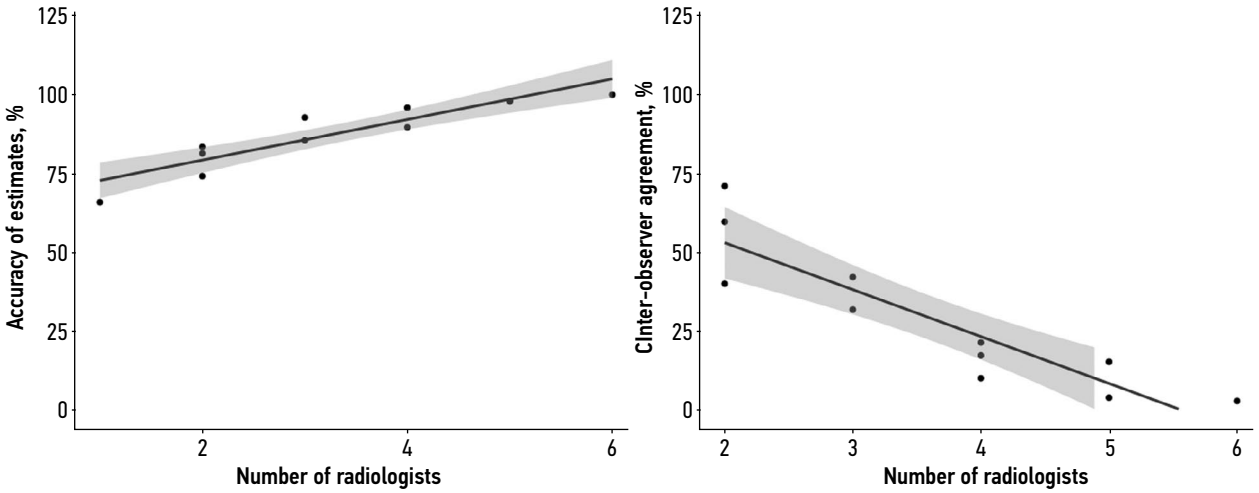

Fig. 1. Accuracy and agreement of estimates as a function of the number of radiologists participating in the primary tagging. The 95% CI is presented in gray. The points correspond to different samples of primary experts. For experiments with two, three, and four experts, three different samples were selected from the initial six radiologists; two various samples were used for five experts.

were made by one expert out of six, 8.1% by two experts, 8.1% by three experts, 19.3% by four experts, and 30.6% by five experts (Fig. 2).

Markup cost

To assess the optimal efficiency of tagging from the standpoint of the rational use of resources, we considered the cost of involving additional experts in the interpretation of CT images. Thus, the improvement in accuracy can be balanced against the increased cost of annotating the studies.

Given that volunteer radiologists participated in tagging the dataset, their work was not paid. Thus, we calculated the cost of tagging in terms of the time spent by the experts. On the average, the primary expert spent 12 min on

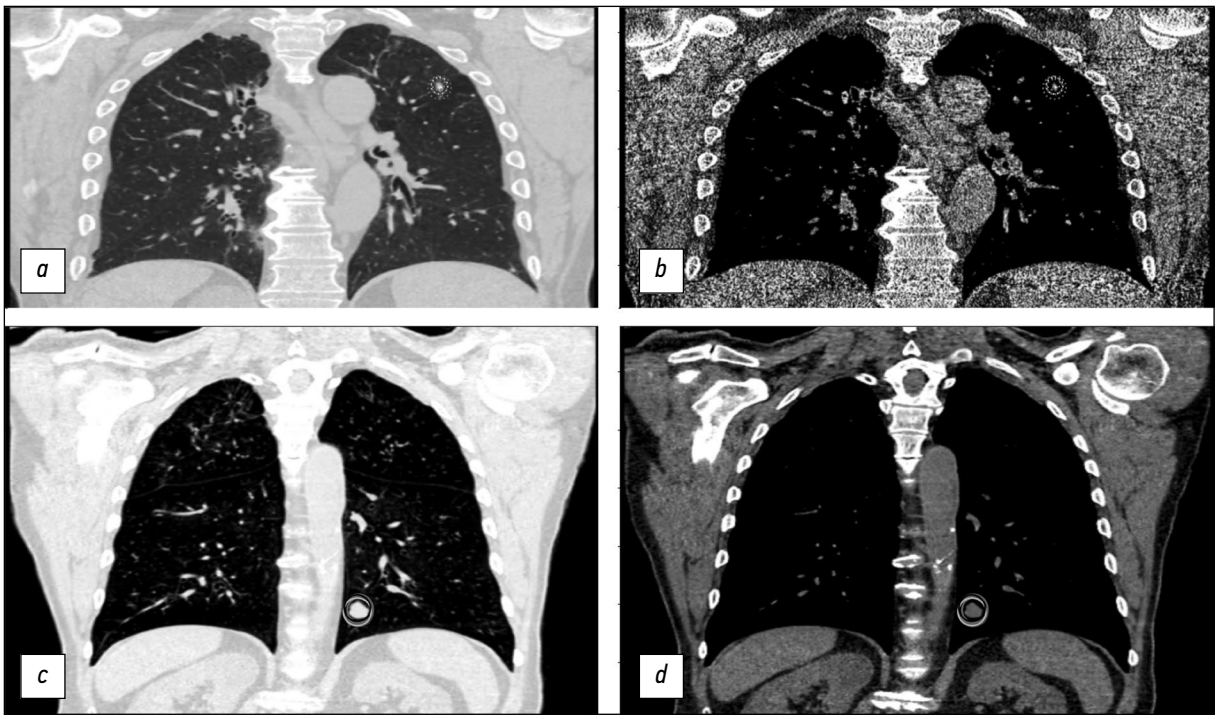

Fig. 2. Examples of CT studies with significant disagreement (*a* and *b*; CTLungCa-500 AN RLADD02000018919, ID RLSDD02000018855) and full consistency (*c* and *d*; CTLungCa-500 AN RLAD42D007-25151, ID RLSD42D007-25151) between experts. The studies are presented in frontal projection in pulmonary (*a* and *c*) and soft tissue (*b* and *d*) modes. The vertical division is 50 mm, and the horizontal division is 100 pixels. The radiologists' marks are presented with different colors: *a* and *b*: the nodule was tagged by five primary experts out of six; four experts classified it as a solid type, and one expert classified it as a semi-solid one. The arbitrator disagreed with their opinion, recognizing the finding as benign calcification; *c* and *d*: all six primary experts and the arbitrator classified the lesion as a potentially malignant solid.

Table 4. Estimated cost of error elimination

Number of primary experts	Number of errors eliminated	Cost, min/error
2	15	129,3
3	19	183,8
4	29	173,9
5	31	212,8
6	33	246,9

the interpretation of one CT image, and the arbitrator spent 4 min. In the present study, the cost of eliminating error C in the studied sample of 97 CT images was calculated as the difference in the average cost of tagging by a given number of primary experts with the involvement of an arbitrator and the cost of tagging by one radiologist without the involvement of an arbitrator divided by the number of errors eliminated (N_{err}):

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

where n is the number of primary experts.

Expert 012+ committed 33 FP and FN errors. Table 4 presents the number of errors eliminated due to attracting additional experts and conducting arbitration and the corresponding cost of eliminating the error. We observed a dependence according to which each new primary expert increased the cost of error elimination by 42.5 ± 10.7 min, excluding one point. The tagging of the dataset by four primary experts with subsequent arbitration was accompanied by a rapid increase in the number of eliminated errors and a decrease in cost (Table 4).

Additional research findings

Given the aspects of the study design, in which each expert interpreted an individual CT scan only once, this study did not assess the intra-observer agreement among individual radiologists. The average value of inter-observer agreement between pairs of specialists was 60.5% ± 5.3%, with a minimum value of 53.1% and a maximum value of 73.0%.

Another way to assess the agreement between primary experts was the analysis of positive findings of each radiologist (Fig. 3). For each representative of the initial cohort, the maximum proportion of detected nodules (37.6% ± 5.4%) corresponded to unique findings that were not recognized by other experts (Fig. 3*a*). Then, in descending order, the findings were approved by one (21.4% ± 2.8%), two (14.0% ± 2.0%), four (9.5% ± 2.3%), three (9.2% ± 1.8%), and five (8.1% ± 3.1%) primary experts. The proportion of

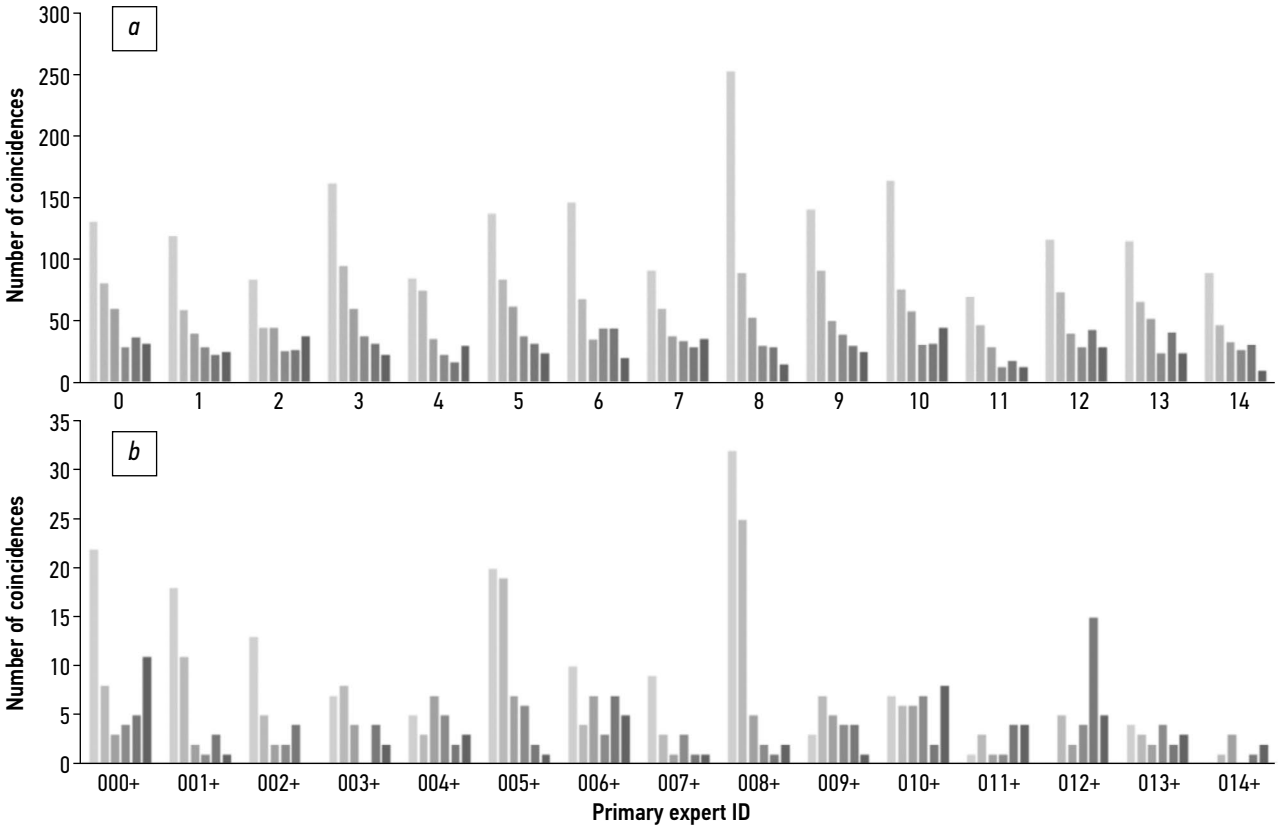

Fig. 3. Agreement between primary experts: *a*. representatives of the initial cohort of 15 radiographers; *b*. replacement radiographers. The data for the expert with ID 000++ are not given due to the small number of lesions annotated. For each radiologist, Column 1 corresponds to the number of lesions tagged uniquely by that specialist (none of the other five experts recognized this finding). The following are columns corresponding to cases where the lesion identified by the radiologist was noted by one, two, three, four, and five other primary experts. The graph disregards the approval of the arbitrator and the differences in the opinion between radiologists about the type of lesion.

unanimously approved findings exceeded 10% for four radiologists from the initial cohort (ID 002, 004, 007, and 010). None of these experts was included in the leading group in terms of Youden's index, which was calculated in accordance with the methodology proposed in this work. Moreover, Radiologist 004 showed the poorest performance in the cohort for this indicator (Table 1). Meanwhile, Radiologist 014, which showed the highest Youden's score in the cohort (0.195), did not stand out among his colleagues in terms of the consistency of positive findings (Fig. 3*a*).

The cohort of radiographers who replaced the initial primary experts had a different distribution of finding agreement (Fig. 3*b*). The maximum proportion of identified nodules (28.9% ± 18.2%) was still represented by unique findings. This result was followed by findings identified simultaneously by two (23.3% ± 11.0%), three (13.3% ± 10.7%), five (13.2% ± 11.9%), six (11.5% ± 9.8%), and four (9.7% ± 7.6%) experts. This cohort had eight radiographers (ID 000+, 004+, 006+, 010+, 011+, 012+, 013+, and 014+), for which the proportion of unanimously approved positive findings exceeded 10%, and the value was above 20% for four of them (ID 000+, 010+, 011+, and 014+). Nevertheless, these indicators may be due to the small number of positive findings in this cohort, which is indirectly evidenced by the high variation in their consistency, expressed in terms of mean values and standard deviations. For example, Expert 014+ participated in the interpretation of CT studies, where other experts identified 63 entities (Table 1). This expert tagged seven nodules, one of which was identified by another expert, three by two experts, one by five experts, and two nodules by six experts (Fig. 3*b*). Furthermore, the expert committed 32 FN errors, thus ignoring approximately 50% of true positive findings. For this cohort, no correlation was registered between the consistency of the positive findings and the expert's Youden's score.

DISCUSSION

Summary of the main research findings

Our results demonstrated that an increase in the number of specialists conducting an independent interpretation of CT studies led to an increase in the accuracy of their estimates, and the level of qualification showed no significant effect on either the consistency of opinions of radiologists or their joint accuracy. Among the factors affecting the inter-observer agreement between the pairs of researchers, a discordance of opinions was observed concerning the presence of lesions in a particular area of the CT scan.

Main research results

No consensus is currently available regarding the recommended number of radiologists to participate in the primary markup and annotation of medical imaging datasets. In general, this number ranges from one [18, 19] to four [20]. Only the work by P.G. Herman and S.J. Hessel addressed this issue; according to their research, the number of error-free descriptions gradually decreases with the increase in the number of specialists providing independent interpretations of studies [5]. Although this finding piques interest, it is of little practical value because the arbitrage model is, in principle, based on the assumption that primary interpretations comprise errors. Moreover, its efficiency increases provided that these errors are different.

The last statement is not always true. In particular, the results of this work indicate that radiologists committing different mistakes does not lead automatically to an increase in the joint accuracy of their conclusions. In an experiment with two specialists who performed the primary interpretation of CT images, the highest level of disagreement was registered in pair 2 (agreement 40.2%), which had also the lowest accuracy of the three considered pairs (74.2% versus 81.4% and 83.5%). In addition, pair 3 showed the highest accuracy value with the maximum agreement (71.1%). Nevertheless, according to the data obtained in this work, a significant negative correlation existed between the agreement of expert assessments and their accuracy (r = –0.78). Thus, at the initial interpretation by two radiographers, the agreement of 57.0% ± 15.6% was noted, with the accuracy of 79.7% ± 4.9%. For five radiographers, these indicators were equal to 9.8% ± 8.1% and 97.9% ± 0.1%, respectively, and this dependence was retained in all the considered variants of dataset tagging (Fig. 1).

According to the results of this study, the optimal combination of accuracy and markup cost can be achieved by an approach involving four primary experts and subsequent arbitration (Table 4). In that case, a rapid increase in the number of eliminated errors was observed in comparison with the tagging by three radiologists, accompanied by a decrease in the time spent on eliminating one error (–9.9 min). The involvement of additional primary experts led to a further increase in the accuracy of interpretations. However, this finding was due to an increase in the cost of eliminating errors by an average of 42.5 ± 10.7 min.

In the present work, when classifying the assessments of primary experts to the categories of FN, TN, FP, and TP, we relied on the assumption that all pulmonary nodules will be tagged on each CT scan. However, the study results indicated that the study participants limited themselves to the five largest pulmonary lesions on CT scans, following the recommendations given to them. Thus, some pulmonary nodules were ignored by individual radiographers, which affected their diagnostic accuracy and the inter-agreement values in expert pairs. Nevertheless, differences in the opinions between primary experts are a desirable outcome when using arbitration because they expand the range of tagged lesions. This condition reduces the proportion of FN findings, even under artificial restrictions on the number of nodules to be tagged. One of the main outcomes of this work is that consensus among several radiographers

is not a prerequisite for proper tagging of datasets. The arbitrators bear the main responsibility because they must correctly interpret all entities noted by the primary experts (Figs. 2*a* and 2*b*).

Research Limitations

The main limitation of this work was the model for determining the ground truth, that is, the findings that should be considered pulmonary nodules. When interpreting CT scans, radiologists lacked access to the clinical, biological, and genomic data of patients. Moreover, the set did not contain two studies that spread out over a period of time, which would have enabled the assessment of the dynamics of development of lesions, for any of the patients. We also proceeded from the assumption that the opinion of the arbitrator is always correct, and we interpreted the disagreements between the primary experts and the arbitrator always in favor of the latter. However, the set presented a number of examples that raised doubts about the reliability of this approach. In particular, 19 pulmonary lesions were tagged by the arbitrator as both benign and malignant. This result is consistent with the results of S.J. Hessel et al., who demonstrated that arbitrators can resolve correctly about 80% of disagreements between primary experts [4].

Another limitation of the work was the inability to assess the reproducibility of the conclusions of individual radiographers. A limited sample was used to achieve the main objectives of the study. For more reliable statistics, the optimal approach would be the bootstrap method. Finally, the assessment of the diagnostic accuracy of the primary experts in the present study relied on the assumption that they would mark all pulmonary nodules. If more than five lesions were observed on the CT scan, this assumption was in conflict with the recommendations for tagging, which can affect the final individual indicators of Se and Sp. To compensate for this methodological limitation, the study authors attempted to assess the consistency in the number of positive findings for each primary examiner approved by two, three, four, and five other radiographers (Fig. 3). However, such an analysis neglected the FN errors, and therefore, its results showed no correlation with the obtained values of Youden's index for each expert. In addition, this study analyzed the results of interpretation of standard dose CT scans. Thus, its findings may not apply to the data obtained from screening studies characterized by the use of low-dose and ultra-low-dose CT protocols.

CONCLUSION

Despite its limitations, this work demonstrated convincingly that an increase in the number of independent primary interpretations can increase their accuracy, if the arbitration is performed. In addition, the qualifications of radiologists are not the decisive factor of the quality of their analysis because according to the results obtained, the joint accuracy of their assessments was independent of individual Youden's indices. The optimal combination of accuracy and cost of tagging was achieved during the initial independent interpretation of CT examinations by four experts. This statement created a theoretical basis for the development of requirements for artificial intelligence algorithms intended for use in the diagnosis of diseases by tagging suspicious structures on CT scans, guiding and attention of radiologists. In addition, the results obtained in this work enable the substantiation of the project model for crowd-tagging of datasets, in which an increase in the number of taggers will lead to a decrease in agreement and a simultaneous increase in the quality of the final product, given arbitration.

ADDITIONAL INFORMATION

Funding. This study was not supported by any external sources of funding.

Conflict of interest. The authors declare that they have no competing interests.

Authors' contribution. All authors made a substantial contribution to the conception of the work, acquisition, analysis, interpretation of data for the work, drafting and revising the work, final approval of the version to be published and agree to be accountable for all aspects of the work. The largest contributions are as follows: N.S.Kulberg – dataset design, conceptualization of the study, preparation and editing of the text of the article; R.V. Reshetnikov – statistical analysis, writing of the manuscript; V.P.Novik – dataset preparation, software development for data processing, statistical analysis; A.B.Elizarov – dataset preparation, software development for data processing; M.A.Gusev – dataset preparation, software development for data processing; V.A.Gombolevskiy – conceptualization of the study, dataset design; A.V.Vladzymyrskyy – conceptualization of the study, editing of the text of the article; S.P.Morozov – dataset design, conceptualization and funding of the study.

Acknowledgments. The authors express their deepest gratitude to Valeria Yurievna Chernina for methodological consultations and to all radiologists who took part in tagging of the dataset.

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

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

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

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

* Corresponding author / Автор, ответственный за переписку