Development of an AI-based Smart Imagery Framing and Truthing (SIFT) system to annotate pulmonary abnormalities with corresponding boundaries based on CT images

Lin Guo[*a], Lingbo Deng[*b], Stefan Jaeger[c], Bin Zheng[d], Qian Xiao[a], Teresa Wu[e], Fulin Cai[e], Fleming Y.M. Lure[d], Li Xia[#a], Weijun Fang[#f]

[a]Shenzhen Zhying Medical Imaging Co., Ltd, Shenzhen, China; [b]Department of Radiology, Peking University Shenzhen Hospital, Shenzhen, China; [c]National Library of Medicine, National Institutes of Health, Bethesda, MD, USA; [d]MS Technologies Corp, Rockville, MD, USA; [e]School of Computing and Augmented Intelligence, Arizona State University, Tempe, AZ, USA; [f]Guangzhou Key Laboratory of Artificial Intelligence for Medical Imaging of Tuberculosis, Department of Radiology, Guangzhou Chest Hospital, Institute of Tuberculosis, Guangzhou Medical University, Guangzhou, China

Lin Guo and Lingbo Deng contributed equally to this work.
[#]Corresponding authors: Li Xia, Email: smileli3254@163.com, and Weijun Fang, Email: fangweijun71@163.com.

ABSTRACT

Clinically significant pulmonary abnormalities, like tuberculosis, nodules, and chronic obstructive pulmonary disease (COPD), hinder lung function. As a non-invasive diagnostic tool, radiology can identify various pulmonary conditions. Artificial intelligence (AI) advancements, particularly deep learning-based applications, have improved the efficiency and accuracy of radiological examinations. However, creating accurate AI models is still challenging due to the requirement for extensive manual annotation. To address this, we propose Smart Imagery Framing and Truthing (SIFT) to assist annotators in labeling and delineating pulmonary abnormalities and corresponding boundaries on CT images. Integrating Mask-RCNN, the system is developed on a training dataset of 9,078 CT images with 166,724 slices and 18,367 regions of interest (ROIs) corresponding to 47 different abnormalities/diseases. The proposed system processes an independent testing set with 2,199 CT images containing 40,493 slices and 4,280 ROIs by predicting both abnormality/disease types and corresponding ROI boundary locations. Two senior radiologists ($\geqslant$ 30 years of radiological experience) delineated each image of training and testing sets in a double-blinded way, and their intersection area functioned as the ground truth. Evaluation metrics include intersection over union (IOU) value calculated based on the ROI area and the ground truth, the area under the ROC curve (AUC), and sensitivity on the slice-level of abnormality label category. For ROI segmentation, the results showed that 91.5% and 36.2% abnormalities had an IOU over 0.6 and 0.7, respectively. Regarding the metric of AUC of classification, there were 97.9%, 80.9%, and 42.6% of abnormalities whose AUC values were above 0.7, 0.8 and 0.9, respectively. For the sensitivity metric, all 47 label categories exceeded 0.8, and 68.1% (32/47) exceeded 0.9. SIFT demonstrates high performance in determining abnormality/disease types with corresponding boundary locations for the ROIs, which can be used to predict training and testing sets to develop AI technologies.

Keywords: CT; artificial intelligence; image annotation; model development; major pulmonary abnormality

1. INTRODUCTION

Developing machine learning and deep learning (ML/DL) for medical applications requires a very large amount of high-quality data (radiographs, genomics, pathological images, etc.) to train and test to reach high accuracy and robustness. High-quality data requires a gold standard to confirm the cases (e.g., biopsy to confirm cancer on chest X-ray and CT) and accurate and consistent annotation (e.g., a specific type of abnormality/disease, its location on the image, and accurate delineation of abnormality/disease boundary). However, such annotation is challenging, requiring the participation of annotators with specific domain knowledge operating annotation tools that are often time-consuming, tedious, inefficient,

Medical Imaging 2025: Computer-Aided Diagnosis, edited by Susan M. Astley, Axel Wismüller, Proc. of SPIE Vol. 13407, 134070M · © 2025 SPIE · 1605-7422 · doi: 10.1117/12.3047186

and lack accuracy [1]. Although some commercial image-based annotation tools have been developed for radiological images, they still cannot meet such demand because annotators need to determine the type of diseases and use a pointing device to follow, draw, and edit the lesion boundaries. This process may deter the annotator from performing such a task. Furthermore, annotation consistency remains a challenge even for identical annotators (intra-observer variance) [2], and different annotators may use different medical terminologies for identical diseases (inter-observer variance) [3].

Recently, DL has demonstrated considerable potential as a tool for abnormality/disease lesion localization in chest CT images. Yan et al. [4] introduced a deep neural network for segmenting COVID-19 infection regions with a dice score of 0.726. Pulmonary nodule localization was also predicted using DL, and an accuracy of 0.923 was finally obtained [5]. An attention-guided DL network was also proposed and tested in a set of 2,160 patients to locate pulmonary embolism, and the result showed that the algorithm could provide localized attention maps for possible pulmonary embolism lesions with an AUC of 0.812 [6]. However, most of these algorithms specialize in a single disease, narrowing their utility in real-world applications. An algorithm capable of segmenting multiple lesions in chest CT images would be a more powerful tool to address training data scarcity and avoid human annotation efforts.

In this paper, the Smart Imagery Framing and Truthing (SIFT) system was developed to address the above issues, aiming to facilitate fully automatic, fast, and accurate ML/DL development. We present the SIFT system's framework and graphic user interface, highlighting its potential benefits in the CT image annotation process for 47 pulmonary abnormalities/diseases through independent validation. The SIFT system possesses the capability to assist annotators in generating high-quality image data labels, proving valuable in supporting the development of ML/DL for medical applications.

2. METHODOLOGY

2.1 Internal Training Dataset and External Testing Dataset

The internal training dataset consists of 9,078 abnormal CT images confirmed by the original radiological report. This dataset had 166,724 slices and 18,367 regions of interest (ROIs), representing 47 different abnormalities. Additionally, an independent set of 2,199 CT images containing 40,493 slices and 4,280 ROIs was used to evaluate the SIFT system's performance in labeling and delineating major pulmonary abnormalities/disease lesions.

2.2 Model Development

Figure 1 illustrates the framework of the proposed SIFT system for lesion-based annotations on CT images. It begins with image studies and predetermined ROIs. Using a disease-guided ROI location approach, relevant areas are identified, followed by ROI-constrained fine-grained segmentation to enhance the accuracy of feature extraction. A step labeled "MOM" (Multi-task, Optimal-recommendation, and Max-predictive Classification and Segmentation) integrates these processes [7], and the results are validated through physician confirmation to ensure clinical reliability. The validated data is then utilized for ML/AI models, which are subsequently tested on new imaging datasets for annotation purposes and performance evaluation.

SIFT was developed based on Mask-RCNN3D, whose framework is similar to the structure of Mask-RCNN and consists of three parts. The first part is the feature extraction Backbone Network modified from Resnet. The basic layers, such as 2D convolutions and Batch Normalization, are replaced with 3D convolutions and Batch Normalization to extract features from 3D CT images. The second part is the Region Proposal Network (RPN), which comprises several layers of 3D convolutions and generates ROIs and their locations near each pixel in the feature maps. The third part is the ROI network, which consists of 3D convolutions and fully connected layers. Mask-RCNN3D is a supervised learning network, and in the study, all training and testing images were labeled and annotated by two senior radiologists (with more than 30 years of radiological experience). Each CT scan in the training data is annotated with pixel-level segmentation of all lesion areas. In the RPN classification, the strategy for selecting anchor box samples is choosing 50% of the samples with the highest intersection over union (IOU) and randomly selecting 50% of samples whose IOU is under 0.3. The training of the ROI network is the same as in Mask-RCNN, where a fixed total number of proposal labels and a maximum number of positive samples are used to update the parameters.

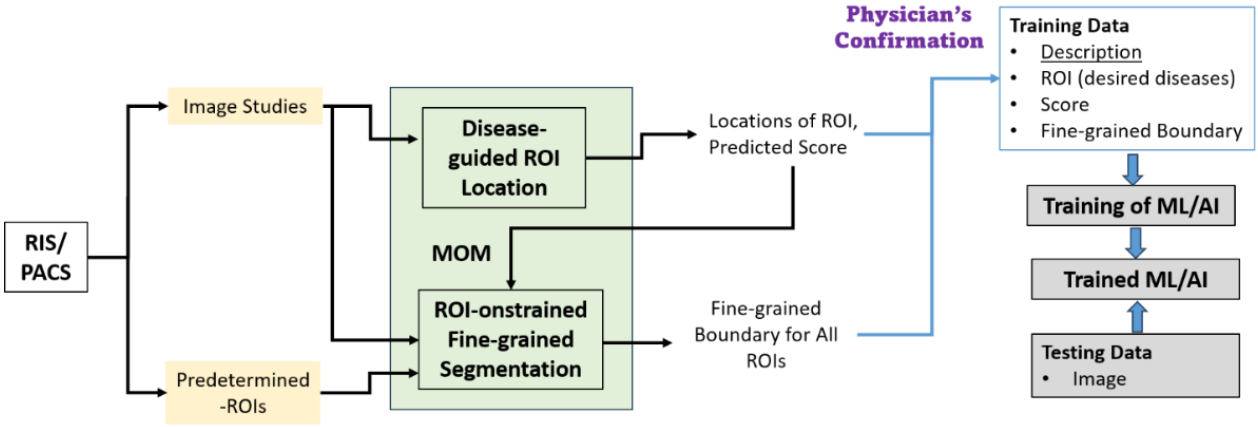


Figure 1. Framework of the SIFT system for lesion-based annotations on CT images.

2.3 Ground Truth (GT) Establishment and Model Performance Evaluation

Two senior radiologists delineated each image of the training and testing sets in a double-blinded way, and their intersection area functioned as the GT for the system segmentation evaluation. In the testing, the SIFT system determined the abnormality types of labeled ROIs and their boundary first, and then we calculated the IOU value based on the ROI area and the GT. The evaluation metric of the area under the ROC curve (AUC) was calculated, and all the evaluations were conducted on the slice level of abnormality classification. The usage of SIFT includes:

a. **SIFT Automatically Predicted Outputs:** All the predicted abnormalities/diseases at different locations with corresponding boundaries of ROIs are displayed on the image. Among these unique abnormalities predicted by SIFT, 90% of disease lexicons comply with Medical Dictionary for Regulatory Activities (MedDRA) UMLS CUI coding, and the rest of 10% comply with Radiopaedia (international collaborative radiology educational resource from Wikipedia).

b. **Output in Excel File:** Consisting of five parts:
Sequence No: Order of all uploaded CT images. It starts with 0.
File_name: It is obtained directly from each file.
Predicted Abnormality/Disease for Each ROI: The SIFT system has predicted all the classes of abnormality at different locations. Their boundary locations are provided at each slice.
Predicted Score for Each ROI: The SIFT system has predicted all the classes of abnormality/disease at different locations. The predicted score (i.e., from 0 to 1) for each ROI is provided.
Location of Boundary for Each ROI: The system has predicted all boundary locations. Each x, y location is grouped by [x, y], and each set is grouped by [ ] and separated by a comma (,).

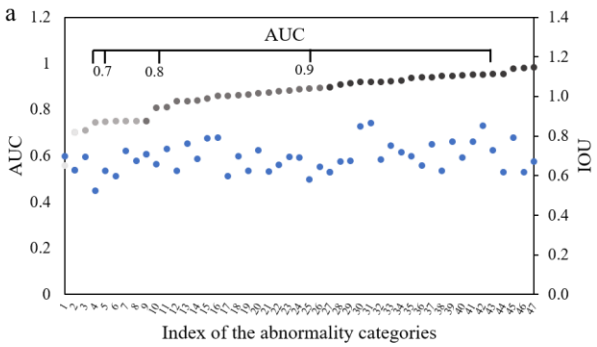
3. RESULTS

In the study, the SIFT system performs well in labeling and delineating 47 different ROIs on CT images, providing confidence scores for detected ROIs.

ROI localization For general object localization, IOU is one of the most commonly used evaluation metrics to measure how close the predicted bounding boxes are to the GT bounding boxes. An IOU of 0.5 often indicates an acceptable threshold across various areas and fields [8], but in medical imaging, a value over 0.3 is often considered sufficiently accurate [9]. It may be due to the complexity and variability of medical images, where precise lesion boundaries of irregular shapes are more challenging to detect and locate. In our study, the IOU values achieved by the proposed SIFT system on the slice-level demonstrate a range of 0.525-0.867, indicating consistent performance across 47 different abnormality categories (Figure 2). An analysis of the results demonstrated that 91.5% (43/47) of abnormalities have an IOU over 0.6, suggesting that the majority of the lesion location predictions align well with the GT annotations. Additionally, 36.2% (17/47) of abnormalities achieve an IOU over 0.7, further revealing the SIFT system's ability to localize abnormalities in many cases on CT images accurately. Three abnormalities/diseases of chronic obstructive pulmonary disease (COPD),

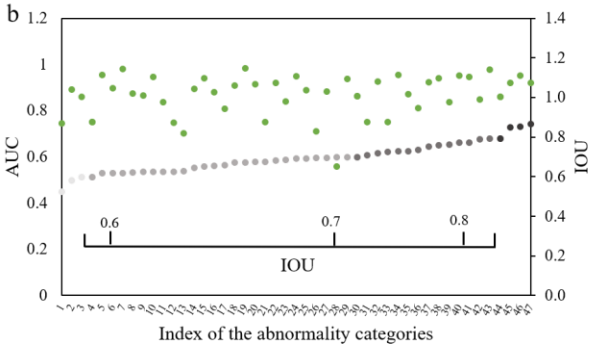
cardiomegaly, and pulmonary edema possessed the highest IOU values; they were 0.848, 0.852, and 0.867, respectively. High IOU values may enhance the credibility of automated diagnostic tools based on this model in real-world clinical settings. Overall, the IOU values appear relatively stable across most categories, indicating the model's robustness in processing various abnormality types. Only a few types show slightly lower performance, likely due to the limited number of CT images available for these cases, leaving room for further improvement in certain specific cases.

Abnormality/Disease Classification Regarding the metric of AUC for classification performance, values ranged from 0.559 to 0.983. Specifically, there were 97.9% (46/47), 80.9% (38/47), and 42.6% (20/47) of abnormalities/diseases whose AUC values were above 0.7, 0.8, and 0.9, respectively. Notably, we observed that eight abnormalities/diseases achieved AUC values equal to or greater than 0.950; they were secondary pulmonary tuberculosis (0.950), bullae (0.951), cardiomegaly (0.951), hepatic cyst (0.955), pericardial effusion (0.956), pleural effusion (0.977), coronary artery atherosclerosis (0.981), and lymph node (0.983) in increasing order. Generally, for abnormalities/diseases with lower AUC values (e.g., below 0.7), the challenges may be attributed to limited training data, overlapping features with other conditions, etc. In our study, it should be noted that among all 47 abnormality/disease categories, only one category, chronic bronchitis of emphysema with infection, has an AUC below 0.7, likely due to the relatively small number of cases. Future improvements could involve collecting more data. As shown in Figure 2, the average AUC was 0.868±0.089, and the average IOU was 0.692±0.074, both having a standard deviation of approximately 10% (10.25% for AUC and 10.71% for IOU), indicating a similar level of variability in these evaluation metrics. This may suggest that the SIFT system demonstrates consistent performance and reliability in multi-disease/abnormality recognition. For the sensitivity, all 47 classification categories exceeded 0.8, ranging from 0.823 to 1.00. Some ML/DL systems have successfully predicted some pulmonary abnormalities/diseases, such as tuberculosis, pneumonia, and pulmonary nodules. However, such ML/DL systems often focus on one specific abnormality/disease binary classification, limiting their utility in general practice where various pulmonary conditions exist. This study involved 47 abnormalities/ diseases, reflecting a broad range of clinically critical pulmonary abnormalities identifiable in radiology. While this study focuses on abnormalities/diseases in chest imaging, the SIFT system could potentially be extended to other imaging modalities or anatomical regions, given its consistent performance.



Index	Abnormality/Disease Name
1	Chronic bronchitis of emphysema with infection
2	Healed rib fracture
3	Pneumothorax
4	Isolated tracheal or bronchial tuberculosis
5	Sclerotic fibrosis
6	Bronchiectasis
7	Metastatic neoplasm
8	Bronchiectasis and infection
9	Disseminated tuberculosis
10	Postoperative changes
11	Atelectasis
12	Fibrogenesis
13	Encapsulated effusion
14	Lung cyst/bronchial cyst
15	Pneumoconiosis
16	Nontuberculous mycobacterial infection
17	Fibrous Focus
18	Mass
19	Pleural thickening
20	Interstitial pneumonia
21	Pleural calcification
22	Rib fracture
23	Pleurisy
24	Bronchitis

Index	Abnormality/Disease Name
25	Subcutaneous emphysema
26	Pneumonia
27	Lymph nodes enlarged
28	Emphysema
29	Viral pneumonia
30	COPD
31	Pulmonary edema
32	Nodule
33	Diffuse panbronchiolitis
34	Interstitial lung disease
35	Pulmonary calcification
36	Interlobar pleurisy
37	Chronic bronchitis & emphysema
38	Arteriosclerosis
39	Hydropneumothorax
40	Secondary pulmonary tuberculosis
41	Bullae
42	Cardiomegaly
43	Hepatic cyst
44	Pericardial effusion
45	Pleural effusion
46	Coronary artery atherosclerosis
47	Lymph node calcification



Index	Abnormality/Disease Name
1	Isolated tracheal or bronchial tuberculosis
2	Subcutaneous emphysema
3	Fibrous Focus
4	Bronchiectasis
5	Pericardial effusion
6	Lymph nodes enlarged
7	Coronary artery atherosclerosis
8	Pleural calcification
9	Pleural thickening
10	Arteriosclerosis
11	Fibrogenesis
12	Sclerotic fibrosis
13	Healed rib fracture
14	Pneumonia
15	Interlobar pleurisy
16	Rib fracture
17	Postoperative changes
18	Emphysema
19	Lymph node calcification
20	Viral pneumonia
21	Bronchiectasis and infection
22	Nodule
23	Lung cyst/bronchial cyst
24	Secondary pulmonary tuberculosis

Index	Abnormality/Disease Name
25	Bronchitis
26	Pneumothorax
27	Pleurisy
28	Chronic bronchitis of emphysema with infection
29	Pulmonary calcification
30	Mass
31	Disseminated tuberculosis
32	Interstitial lung disease
33	Metastatic neoplasm
34	Hepatic cyst
35	Interstitial pneumonia
36	Atelectasis
37	Diffuse panbronchiolitis
38	Chronic bronchitis & emphysema
39	Encapsulated effusion
40	Bullae
41	Hydropneumothorax
42	Pneumoconiosis
43	Pleural effusion
44	Nontuberculous mycobacterial infection
45	COPD
46	Cardiomegaly
47	Lymph node calcification

Figure 2. Testing results of AUC and IOU values of 47 abnormalities/diseases. (a) AUC values are arranged in ascending order, with the corresponding IOU value (the blue dot) of each type of abnormality; (b) IOU values are arranged in ascending order, with the corresponding AUC value (the green dot) of each type of abnormality.

As shown in Table 1, three major clinically significant pulmonary abnormalities/diseases are presented: tuberculosis (both active and inactive), pulmonary nodules, and abnormalities related to COPD. These conditions are also the most common abnormalities/diseases in the testing set. Tuberculosis is one of the most deadly infectious diseases, while pulmonary nodules are considered a great contributor to lung cancer. The above two are popular research subjects in AI medical imaging. COPD remains a leading cause of mortality worldwide, and pulmonary emphysema, together with chronic bronchitis, is a part of COPD. However, few reported studies on AI concerning COPD have been reported. The study provides COPD-related abnormality labels, such as pulmonary emphysema and chronic bronchitis with emphysema. Specific disease diagnoses of COPD could be further made based on identifying these abnormalities, together with pulmonary function test results. This approach simulates the radiologist's interpretation process, incorporating abnormality descriptions into disease diagnosis, leading to more appropriate deployment in clinical practice. Among these three categories, COPD-related abnormalities/diseases achieved the highest sensitivity (0.910 for chronic bronchitis and emphysema) and the highest IOU (0.848 for COPD), and the AUC values for all COPD-related abnormalities/diseases are greater than 0.9. Pulmonary nodules achieved an AUC of 0.921, further reflecting the SIFT system's robustness. For tuberculosis, especially secondary pulmonary tuberculosis, the system possesses a high AUC value of 0.950 and a high sensitivity of 0.899, showing the SIFT system's ability to recognize this infectious disease. It is reported that most active pulmonary tuberculosis belongs to secondary pulmonary tuberculosis [10]; therefore, effectively and accurately locating the lesions is useful for the prevention and control of tuberculosis. Figure 3 displays three examples of the SIFT system's graphical user interface applied to CT images of these three categories. These visual results highlight the SIFT system's capability to locate and classify abnormalities precisely.

Table 1. Annotation results of the three most common types of pulmonary abnormalities/diseases

No.	Abnormality/Disease Name		IOU	AUC	Sensitivity
1	Active tuberculosis	Secondary pulmonary tuberculosis	0.691	0.950	0.899
		Disseminated tuberculosis	0.710	0.752	0.890
2	Inactive tuberculosis	N/A	0.626	0.747	0.871
3	Pulmonary nodules	N/A	0.682	0.921	0.893
4	COPD-related	Emphysema	0.672	0.909	0.899
		Chronic bronchitis and emphysema	0.760	0.941	0.910
		COPD	0.848	0.919	0.860

Figure 3. Examples of the SIFT graphic user interface: secondary tuberculosis (left), a nodule (middle), and emphysema (right) have been predicted by the SIFT system, and their boundary locations are provided.

4. CONCLUSIONS

We developed and evaluated a novel AI-based SIFT system that demonstrates high efficiency and accuracy in annotating multiple pulmonary abnormalities/diseases on CT images. This system would benefit annotators in determining the abnormality/disease types, their boundary coordinates, and confidence levels for CT images. The SIFT system's performance in recognizing diverse pulmonary conditions supports its potential application in automating large-scale image annotation tasks, reducing manual workload, and enhancing annotation consistency. Moreover, the SIFT system could potentially be extended to other imaging modalities or anatomical regions, given its consistent performance.

ACKNOWLEDGEMENTS

This study has received funding from the Shenzhen Science and Technology Program [Grant No.: KQTD2017033110081833; JCYJ20220531093817040], the Guangzhou Science and Technology Planning Project [Grant No.: 2023A03J0536; 2024A03J0583], and the Inner Mongolia Autonomous Region Science and Technology Program Project [Grant No.: 2024SGGZ059]. This research work was supported in part by the Lister Hill National Center for Biomedical Communications of the National Library of Medicine (NLM), National Institutes of Health.

REFERENCES

[1] Tajbakhsh N, Jeyaseelan L, Li Q, et al. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation[J]. Medical image analysis, 2020, 63: 101693.
[2] Abujudeh H H, Boland G W, Kaewlai R, et al. Abdominal and pelvic computed tomography (CT) interpretation: discrepancy rates among experienced radiologists[J]. European radiology, 2010, 20: 1952-1957.
[3] Bello S M, Shimoyama M, Mitraka E, et al. Disease Ontology: improving and unifying disease annotations across species[J]. Disease models & mechanisms, 2018, 11(3): dmm032839.
[4] Yan Q, Wang B, Gong D, et al. COVID-19 chest CT image segmentation--a deep convolutional neural network solution[J]. arXiv preprint arXiv:2004.10987, 2020.
[5] Ewaidat H A, Brag Y E. Identification of lung nodules CT scan using YOLOv5 based on convolution neural network[J]. arXiv preprint arXiv:2301.02166, 2022.
[6] Shi L, Rajan D, Abedin S, et al. Automatic diagnosis of pulmonary embolism using an attention-guided framework: A large-scale study[C]//Medical imaging with deep learning. PMLR, 2020: 743-754.
[7] Guo L, Hong K, Xiao Q, et al. Developing and assessing an AI-based multi-task prediction system to assist radiologists detecting lung diseases in reading chest x-ray images[C]//Medical Imaging 2023: Image Perception, Observer Performance, and Technology Assessment. SPIE, 2023, 12467: 73-90.
[8] Padilla R, Passos W L, Dias T L B, et al. A comparative analysis of object detection metrics with a companion open-source toolkit[J]. Electronics, 2021, 10(3): 279.
[9] Feher B, Kuchler U, Schwendicke F, et al. Emulating clinical diagnostic reasoning for jaw cysts with machine learning[J]. Diagnostics, 2022, 12(8): 1968.
[10] Wang S H, Satapathy S C, Zhou Q, et al. Secondary pulmonary tuberculosis identification via pseudo-Zernike moment and deep stacked sparse autoencoder[J]. Journal of Grid Computing, 2022, 20: 1-16.