Generating Chinese Radiology Reports from X-Ray Images: A Public Dataset and an X-ray-to-Reports Generation Method

Wen Tang[1], Chenhao Pei[1], Pengxin Yu[1], Huan Zhang[1], Xiangde Min[2], Cancan Chen[1], Han Kang[1], Weixin Xu[1], and Rongguo Zhang[1,3(✉)]

[1] Infervision Medical Technology Co., Ltd., Beijing, China
zrongguo@infervision.com
[2] Department of Radiology, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, No. 1095 Jie Fang Avenue, Wuhan 430030, Hankou, People's Republic of China
[3] Academy for Multidisciplinary Studies, Beijing National Center for Applied Mathematics, Capital Normal University, Beijing 10048, China
zrongguo@cnu.edu.cn

Abstract. Deep learning methods have revolutionized medical image analysis, enabling tasks such as lesion classification, segmentation, and detection. However, these methods rely on annotations, posing a burden on healthcare professionals. In contrast, medical reports contain valuable information, leading to the emergence of Medical Reports Generation from Medical Images (MRGMI). Despite advancements, MRGMI predominantly focuses on English reports, lacking solutions for other languages. To address this and to generate responsible Chinese MRGMI model, we present a Chinese MRGMI dataset of over 40,000 Xray-image-report pairs, covering diverse diseases. We further provide 500 graph-node annotations of the reports and propose the CN-RadGraph model, extracting graph nodes from reports to, in a clinical-responsible way, evaluate our MRGMI model: Chinese X-ray-to-Reports Generation (CN-X2RG) model. Considering linguistic disparities, we enhance the SOTA method with prompt training, graph-based augmentation, and sentence shuffling. Our CN-X2RG model shows significant improvements over baselines. The dataset and code are publicly available, fostering clinical-responsible research and development.

Keywords: Radiology Reports Generation · Dataset · Xray

1 Introduction

Deep learning in medical image analysis has shown great success in tasks like lesion classification, organ segmentation, and lesion detection. However, its

First authors (W. Tang C. Pei—Are with the same degree of contribution, they are the co-first authors).

M. E. Celebi et al. (Eds.): MICCAI 2023 Workshops, LNCS 14393, pp. 79–88, 2023.
https://doi.org/10.1007/978-3-031-47401-9_8

reliance on annotated medical images poses a burden on healthcare workers. Alternatively, leveraging comprehensive medical reports can aid in Medical Reports Generation from Medical Images [9] (MRGMI), reducing reliance on annotations and maximizing clinical data for automated analysis.

While significant advancements have been made in MRGMI, the prevailing focus of most methods and datasets [4,10] lies on English-language reports. However, the critical need arises for MRGMI models and datasets in languages other than English due to the linguistic diversity and variations in report styles within the medical field. The translation of medical terminologies and the precise capture of nuanced language present substantial challenges, especially given the impact of report data on MRGMI models and the model's responsibility in medical practice. Therefore, for contexts like medical diagnosis, where errors are unacceptable, access to native medical report data and corresponding models becomes indispensable. Additionally, the development of comprehensive evaluation criteria is equally vital, encompassing not only the assessment of report generation similarity but also a heightened focus on key diseases within anatomies. High-quality evaluation metrics can contribute to designing and obtaining more reliable and secure MRGMI models.

To address the challenges above, this paper presents the Chinese Chest X-ray (CN-CXR) dataset, comprising 46,301 X-ray image-report pairs, covering a wide range of medical scenarios and diseases detectable through X-rays. Additionally, we provide 500 graph-node annotations for the corresponding Chinese reports. Leveraging this dataset, we propose the CN-RadGraph model, which extracts meaningful graph nodes from medical reports, serving as a crucial component in evaluating the performance of our MRGMI model, Chinese X-ray-to-Reports Generation (CN-X2RG). Both datasets provide the possibility to create responsible-Chinese-MRGMI. Thus, we also introduce enhancements to bridge the gap between English-based methods and the unique linguistic characteristics of the Chinese language. Extensive experiments demonstrate significant improvements over existing baselines, highlighting the effectiveness, relevance and clinical responsibility of our proposed enhancements.

2 Related Work

The task of MRGMI [9], has gained significant attention due to advancements in deep learning, natural language processing (NLP), and multimodal learning. In this review, we highlight key works that have contributed to the MRGMI field, and analyze the potential for enhancing their clinical responsibilities.

Datasets. The availability of suitable datasets plays a crucial role in advancing MRGMI research. Two widely used public datasets in MRGMI are the MIMIC-CXR [10] dataset and the IU X-Ray [4] dataset. However, these datasets only provide valuable resources for English-based MRGMI research, the availability of language-specific datasets is essential to cater to different linguistic contexts. For instance, the CX-CHR [11] dataset, a proprietary internal dataset, addresses the need for Chinese MRGMI research. In parallel efforts, Wang et al.

[16] have attempted to address the need for Chinese MRGMI research by translating English datasets into Chinese using ChatGPT. However, it should be noted that even ChatGPT struggles to precisely translate medical terminology, leading to a considerable gap between the translated reports and the authentic ones. The language of authentic reports is more written and standardized, which facilitates quick reading and comprehension. Therefore, it is irresponsible for clinical application to obtain data only by using language translation.

Considering the limitation of current Chinese datasets and to ensure responsibility of MRGMI model, we have collected a new dataset called CN-CXR. This dataset includes 46,301 X-ray images and their corresponding authentic reports, specifically focusing on the medical findings observed in the X-ray images. By providing a large-scale Chinese MRGMI dataset, we aim to facilitate research in this domain and encourage the development of language-specific MRGMI models.

Auxiliary Models. Auxiliary models play a vital role in enhancing the performance of MRGMI models. One notable auxiliary model is RadGraph [7], which extracts entities and relations from medical reports, thereby providing valuable structured information. By incorporating RadGraph into the MRGMI pipeline, researchers can benefit from its ability to provide semantic information during model training [18], as well as evaluate the performance of MRGMI models [8]. Both provide semantic specifications for model-generated reports to enforce accountability. There are also other model such as Clinical-BERT [1] and RadCLIP [5]. However, it is worth noting that most of the existing auxiliary models are primarily designed for English-based datasets, highlighting the need for language-specific auxiliary models for different linguistic contexts.

To address the language-specific nature of MRGMI, we extend the RadGraph approach to the Chinese domain and propose the CN-RadGraph model. Following a similar annotation process, we annotate entities and relations from 500 samples of our CN-CXR dataset. The CN-RadGraph model serves as a valuable resource for Chinese MRGMI research, offering structured information and evaluation capabilities specific to the Chinese language. This reinforces the clinical responsibility of our model.

MRGMI Models. Recent advancements in transformer models and multimodal learning have paved the way for the development of various MRGMI models. For instance, [3] and [13] leverage memory-driven Transformers to generate radiology reports. [2] utilizes a memory bank to enforce consistency between input image features, while [14] employs uncertainty and Kullback-Leibler similarity to maintain consistency between image and report features. Similarly, [20] incorporates a weakly supervised contrastive loss, and [8] utilizes contrastive learning and matching techniques to improve MRGMI performance. Furthermore, auxiliary models can help enhance report generation. Both [18] and [21] use RadGraph model to provide knowledge to their models.

In this paper, we present improvements to existing MRGMI model for our CN-CXR dataset. We utilize the MedCLIP [19] pretrained model as a prompt training feature and employ a memory-driven Transformer architecture for

radiology report generation. The CN-RadGraph model assists by providing a classification auxiliary loss, guaranteeing semantic consistency to ensure clinical responsibility. To address limited training data, we propose a novel report data augmentation method and a data oversampling method, increasing dataset diversity. These enhancements enable our CN-X2RG model as a responsible AI to generate more accurate and comprehensive medical reports for the Chinese domain.

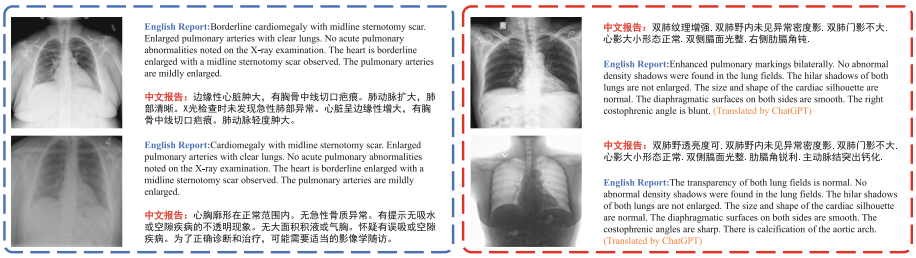

Fig. 1. Illustration of the Chinese(Right)(**The proposed dataset**) and English(Left) reports. The blue box [16] indicates that the original report is in English, and the red indicates that the original report is in Chinese. Using ChatGPT [6] as a translation tool.

3 Datasets

3.1 CN-CXR Dataset

The proposed method is evaluated on our Chinese X-ray reports dataset, CN-CXR, which comprises a total of 46,301 X-ray images along with their corresponding Chinese reports. The dataset was collected from one hospital in China and spans the years 2012 to 2021. In terms of patient demographics, the dataset includes 27,949 male patients, 23,109 female patients, and 923 patients with unknown sex. The age distribution of the patients is reported as 44.18 ± 22.34 years. Table 1 presents the anatomy with lesion extracted using our CN-RadGraph. The table provides an overview of the approximate distribution of some diseases in the dataset. It offers valuable insights into the prevalence and occurrence of various diseases within the extracted anatomical structures. Figure 1 illustrates examples of Chinese reports and their corresponding English translations, as well as English reports and their corresponding Chinese translations. It is evident that the language descriptions and styles differ significantly between the two. Such differences can result in substantial domain shifts, potentially causing a model trained on English reports to fail when generating Chinese reports with translations, and vice versa. Therefore, such a clinically relevant CN-CXR dataset is essential to obtain responsible Chinese MRGMI models.

Table 1. The 48 keywords for *anatomy* and *lesion* selected from the Chinese reports. A *lesion* present on an *anatomy* counts both the *anatomy* and the *lesion* once.

Anatomy Keywords				Lesion Keywords			
	train	valid	test		train	valid	test
Chest	3765	1000	1897	Postoperative	3125	599	1246
Hilum	2803	1309	2271	Gas	783	132	283
Heart	3431	890	1686	Shadow	13911	4210	7798
Diaphragmatic	3096	981	1718	Texture	19774	2578	6082
Costophrenic Angle	5670	1645	3038	Calcification	1431	299	608
Lung	35245	7534	15167	Insert	2375	272	662
Aorta	1888	485	925	No Lung Texture	762	388	676
Rib	2488	299	715	Effusion	743	338	512
Mediastinum	870	311	574	Transparency	1557	606	1067
Pleura	867	327	559	Fracture	131	22	56
Vertebrae	703	239	407	Prominence	1441	361	657
Aortic Knot	2387	534	1038	Side Bend	429	188	299
Aortic Arch	73	13	38	Intubation	132	4	18
Fissures	141	55	89	Deformed	47	25	41
Bone	179	35	73	Drainage	100	29	63
Trachea	313	41	123	Emphysema	62	26	48
Neck	165	29	54	Atelectasis	9	4	6
Horizontal Split	40	24	26	Flame	6	3	4
Bronchi	57	22	36	Torque	953	265	500
Bowel	204	11	36	Blur	265	92	201
Diaphragm	274	90	180	Disappeared	23	12	12
Pulmonary artery	183	79	142	Weaken	123	80	126
Belly	72	10	26	Enhanced	4734	1343	2539
Yessels	6	1	5	Shift	431	153	294

3.2 CN-RadGraph Dataset

To create the CN-RadGraph dataset for our model, we employed a sampling strategy from the CN-CXR dataset. A total of 500 samples were selected based on the length of the reports. The sampling process involved randomly choosing 100 samples, as well as 100 samples from both the top 10% longest reports and the top 10% shortest reports. Additionally, 200 samples were randomly selected from the remaining reports. The resulting 500 samples were then annotated according to the following:

Entities: We define entities as continuous spans of text that can consist of one or more adjacent words. In our schema, entities revolve around two main concepts: Anatomy and Observation. We categorize observations into three types, resulting in four entities in our schema: Anatomy, Disease-Positive-Observation (DPO), Disease-Negative-Observation (DNO), and Neutral-Observation (NO). Anatomy represents anatomical body parts mentioned in the radiology report, such as "lung". Observations encompass words associated with visual features, identifiable pathophysiologic processes, or diagnostic disease classifications. For

example, a DPO could be "effusion" or more general phrases like "increased". A DNO could be "normal" or "no abnormality". A NO could be "shadow", which is always connected to DPO and DNO.

Relations: Relations are directed edges connecting two entities in our schema. We utilize two types of relations: Located At and Modify. Located At (Observation, Anatomy) represents a relationship between an Observation entity and an Anatomy entity, indicating that the Observation is related to the Anatomy. Although Located At often refers to location, it can also describe other relations between an Observation and an Anatomy. Modify (Observation, Observation) or (Anatomy, Anatomy) represents a relation between two Observation entities or two Anatomy entities, signifying that the first entity modifies the scope or quantifies the degree of the second entity. Furthermore, to ensure accurate labeling of diseased or non-diseased conditions, the Observation determining the disease appearance is always placed as the outermost node. Supplementary illustrates an example of a report annotated according to our schema, along with the resulting graph representation.

In CN-RadGraph dataset, we annotate 4 types entities and 2 types relations as shown in Table 2. This carefully curated CN-RadGraph dataset serves as the training and evaluation data for our CN-RadGraph model, enabling accurate and context-aware analysis of radiology reports, which ensures the clinical responsibility of our model.

4 Method

We first formulate the problem of MRGMI mathematically. Formally, we have access to labeled dataset contained radiogy images and chinese reports, denoted by $(X, Y) = \{(x_j, y_j)\}_{j=1}^{N}$, where N is the number of data points. We encode the radiogy images x_j to patch features squence, denoted by $x_{Pj} = \{x_1^{pj}, x_2^{pj}, \ldots, x_K^{pj}\}$, where K is the length of the patch features from vision extractors and embed the corresponding report y_j to the squence, denoted by $y_{Rj} = \{y_1^{rj}, y_2^{rj}, \ldots, y_M^{rj}\}, y_m^{rj} \in \mathbb{T}$, where y_M^{rj} are the generated tokens, M is the length of the generated tokens and $\mathbb{T}$ is the token library which contains all possible tokens. To learn a reports generation model to predict the reports squence of target image sequence x_{pj}, i.e., $\hat{y}_{rj}$, we propose prompt-based vision feature extraction which combine the prompt module with vision encoder and then employ memory-driven Transformer. Note that for simplicity we denote the proposed framework as CN-X2RG, the Chinese X-ray-to-Reports Generation.

The framework of the proposed method is illustrated in Fig. 2. We first extract the radiology images feature by the proposed Prompt-based Vision Feature Extraction which can fuse the features from prompt module and vision encoder and then transfer the extracted feature to patch features for generation. To achieve more accurate and responsible reports, we employ memory-driven Transformer architecture inspired by the work of [3]. To enhance the representation of the feature, we introduce an extra classification task. The details of the

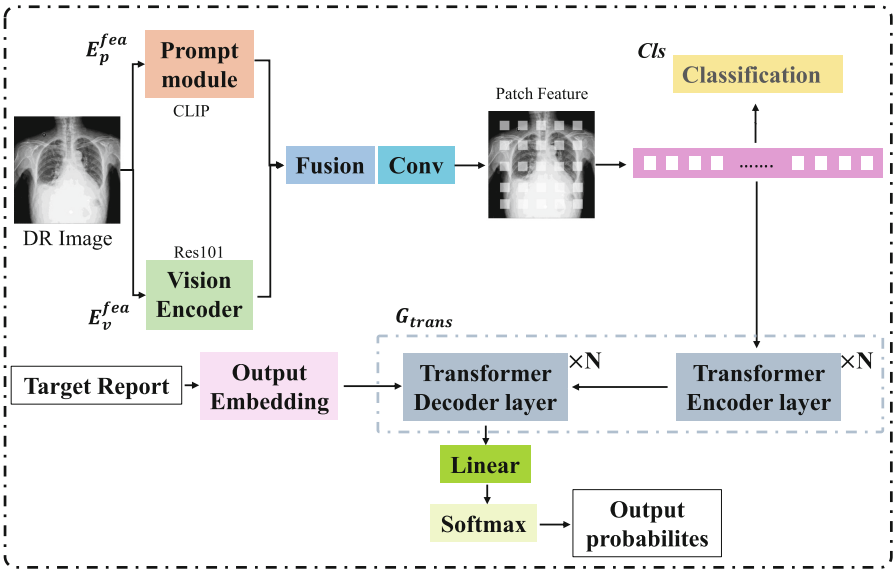

Fig. 2. Overview of the proposed framework, which contains four modules, including prompt module (E_p^{fea}), vision encoder(E_v^{fea}), classification module (Cls) and transformer-based language model (G_{trans}). More specifically, E_p^{fea} and E_v^{fea} are respectively aimed to encoder the learned knowledges and vision features from Xray images. Cls is a auxiliary classification task. G_{trans} performs on a fusion feature space to generate radiology reports.

above, the total loss function and the architecture of proposed network can be found in the supplementary material.

Table 2. Annotation statistics of the CN-RadGraph and statistics of the results obtained by CN-RadGraph on the CN-CXR dataset

Data	Anatomy	DNO	DPO	NO	Locate at	Modify
Annotations (425 training)	5,905	1,433	2,932	674	3,400	4,474
Annotations (75 validation)	1,029	243	515	116	591	772
Results (32,410 training)	366,699	142,411	125,583	57,853	209,731	279,770
Results (4,160 validation)	55,566	14,292	25,980	6,867	30,824	43,180
Results (9,731 testing)	123,142	36,844	52,038	16,584	68,714	94,922

5 Results and Analysis

In this section, we first introduce the dataset, pre-processing and evaluation indicators used for experiments in Sect. 5.1. Then, in Sect. 5.2, we compare the proposed CN-X2RG to other state-of-the-art methods. Finally, we analyze the effectiveness of the modules in the proposed method using two ablation studies which can be found in supplementary material.

5.1 Pre-processing and Evaluation

Pre-processing. For CN-CXR dataset, we resized all training images into resolution of 256×256 pixel, and cropped them into an ROI with 224×224 pixel. All images were normalized with the z-score method, which is conducive to the convergence of the network at the training stage. We folllow the method of R2Gen [3] to split the dataset into train/validation/test set by 7:1:2. Specifically, 32410 images for training, 4160 images for validating and 9731 images for testing. And we used pkuseg [12] and jieba [17] for Chinese word segmentation.

During the training stage, to address the class imbalance issue among keywords, we employed an oversampling technique for positive samples with a keyword count smaller than 5,000. A maximum sampling rate of 10 times was set, and a maximum sampling number of 5,000 was imposed to ensure balanced representation across the classes.

Evaluation Indicators. To evaluate the generation accuracies, we used six metrics, i.e., the $BLEU_1$, $BLEU_2$, $BLEU_3$, $BLEU_4$ [15], $METEOR$ and $ROUGE_L$.

Due to the potential proximity of sentences with opposite semantics when evaluated using $BLEU$ scores, we employ the CN-RadGraph model, which is clinical responsibility due to semantic sensitivity, to extract crucial semantic graph for evaluating model performance. For accurate assessment, we consider true positives when both lesion and anatomy in graphs match between the label and prediction. Conversely, false positives are counted when there are incorrect lesion or anatomy predictions, and false negatives are counted when lesions are missing. Based on this definition, evaluation metrics such as recall, precision, and F1-score can be utilized for assessing lesion and anatomy keywords, as demonstrated in the supplementary material.

5.2 Performance and Comparisons

CN-X2RG Model. We did the comparison study on the CN-CXR dataset for Chinese reports generation. The results are presented in Supplementary. CN-X2RG achieved the best performances in $BLEU_4$, $METEOR$ and $ROUGE_L$ values among all the methods. Specifically, CN-X2RG obtained higher $BLEU_4$, $METEOR$ and $ROUGE_L$ scores than R2GenCMN, with a margin about 1.0%, 0.66% and 0.55% than R2GenCMN, respectively. Compared to R2Gen, R2GenCMN obtained better results. The reason could be that R2GenCMN use cross-modal mapping to facilitate radiology report generation, which was more delicate than R2Gen. However, R2GenCMN consumes a significant amount of graphics memory, which is why we chose R2Gen as the base model.

CN-RadGraph Model. To assess the performance of our CN-RadGraph model, we conducted validation using a subset of the CN-CXR dataset called the CN-RadGraph validation dataset. This dataset comprises 75 Chinese X-ray

reports specifically selected for evaluation purposes. We evaluated the model's performance based on mean precision, mean recall, and mean F1-score for affinity and relation. The impressive results obtained from the CN-RadGraph model are presented in Supplementary, highlighting its strong performance in accurately capturing and analyzing the relationships within the radiology reports.

References

1. Alsentzer, E., et al.: Publicly available clinical bert embeddings. arXiv preprint arXiv:1904.03323 (2019)
2. Chen, Z., Shen, Y., Song, Y., Wan, X.: Cross-modal memory networks for radiology report generation. arXiv preprint arXiv:2204.13258 (2022)
3. Chen, Z., Song, Y., Chang, T.H., Wan, X.: Generating radiology reports via memory-driven transformer. arXiv preprint arXiv:2010.16056 (2020)
4. Demner-Fushman, D., et al.: Preparing a collection of radiology examinations for distribution and retrieval. J. Am. Med. Inform. Assoc. **23**(2), 304–310 (2016)
5. Endo, M., Krishnan, R., Krishna, V., Ng, A.Y., Rajpurkar, P.: Retrieval-based chest x-ray report generation using a pre-trained contrastive language-image model. In: Machine Learning for Health, pp. 209–219. PMLR (2021)
6. Floridi, L., Chiriatti, M.: Gpt-3: Its nature, scope, limits, and consequences. Mind. Mach. **30**, 681–694 (2020)
7. Jain, S., et al.: Radgraph: extracting clinical entities and relations from radiology reports. arXiv preprint arXiv:2106.14463 (2021)
8. Jeong, J., et al.: Multimodal image-text matching improves retrieval-based chest x-ray report generation. arXiv preprint arXiv:2303.17579 (2023)
9. Jing, B., Xie, P., Xing, E.: On the automatic generation of medical imaging reports. arXiv preprint arXiv:1711.08195 (2017)
10. Johnson, A.E., et al.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Sci. Data **6**(1), 317 (2019)
11. Li, Y., Liang, X., Hu, Z., Xing, E.P.: Hybrid retrieval-generation reinforced agent for medical image report generation. In: Advances in Neural Information Processing Systems 31 (2018)
12. Luo, R., Xu, J., Zhang, Y., Zhang, Z., Ren, X., Sun, X.: Pkuseg: a toolkit for multi-domain chinese word segmentation. arXiv preprint arXiv:1906.11455 (2019)
13. Miura, Y., Zhang, Y., Tsai, E.B., Langlotz, C.P., Jurafsky, D.: Improving factual completeness and consistency of image-to-text radiology report generation. arXiv preprint arXiv:2010.10042 (2020)
14. Najdenkoska, I., Zhen, X., Worring, M., Shao, L.: Uncertainty-aware report generation for chest x-rays by variational topic inference. Med. Image Anal. **82**, 102603 (2022)
15. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
16. Rongsheng, W., Yaofei, D., Junrong, L., Patrick, P., Tao, T.: Xrayglm: the first chinese medical multimodal model that chest radiographs summarization. https://github.com/WangRongsheng/XrayGLM (2023)
17. Sun, J.: Jieba chinese word segmentation tool (2012)

18. Wang, Z., Tang, M., Wang, L., Li, X., Zhou, L.: A medical semantic-assisted transformer for radiographic report generation. In: Medical Image Computing and Computer Assisted Intervention-MICCAI 2022: 25th International Conference, Singapore, 18–22 September 2022, Proceedings, Part III. pp. 655–664. Springer (2022). https://doi.org/10.1007/978-3-031-16437-8_63
19. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: contrastive learning from unpaired medical images and text. arXiv preprint arXiv:2210.10163 (2022)
20. Yan, A., et al.: Weakly supervised contrastive learning for chest x-ray report generation. arXiv preprint arXiv:2109.12242 (2021)
21. Yang, S., Wu, X., Ge, S., Zhou, S.K., Xiao, L.: Knowledge matters: chest radiology report generation with general and specific knowledge. Med. Image Anal. **80**, 102510 (2022)