LUCAS: LUng CAncer Screening with Multimodal Biomarkers

Laura Daza[1](✉), Angela Castillo[1], María Escobar[1], Sergio Valencia[2], Bibiana Pinzón[2], and Pablo Arbeláez[1]

[1] Center for Research and Formation in Artificial Intelligence, Universidad de los Andes, Bogotá, Colombia
{la.daza10, a.castillo13, mc.escobar11, pa.arbelaez}@uniandes.edu.co
[2] Fundación Santa Fe de Bogotá, Bogotá, Colombia

Abstract. We present the LUng CAncer Screening (LUCAS) Dataset for evaluating lung cancer diagnosis with both imaging and clinical biomarkers in a realistic screening setting. We extract key information from anonymized clinical records and radiology reports, and we use it as a natural complement to low-dose chest CT scans of patients. We formulate the task as a detection problem and we develop a deep learning baseline to serve as a future reference of algorithmic performance. Our results provide solid empirical evidence for the difficulty of the task in the LUCAS Dataset and for the interest of including multimodal biomarkers in the analysis. All the resources of the LUCAS Dataset are publicly available.

Keywords: Early Lung Cancer Diagnosis · Multimodal Biomarkers · Multimodal Dataset · Lung Nodules

1 Introduction

Lung cancer is the second most common type of cancer in the world [11]. Its high incidence and multiple risk factors make it a critical worldwide health problem and the focus of a considerable amount of research. In the last decade, constant progress in the development of pharmaceutical molecules and treatments for lung cancer [8] have ensured that, if detected early, the prognosis is relatively benign and the survival rate high. For instance, recent studies incorporate genetic analysis for cancer progress examination [1]. The aim was to process patient's DNA to identify the recurrence of lung cancer and find the correlation of lung nodules volume with circulating tumor DNA.

However, despite these advances, lung cancer is still the leading cause of death by cancer in the world, surpassing the combined casualties of the next three types of cancer. Every year, around 2.09 million new cases are detected (11.6% of total cases [11]) and nearly 1.76 million people die because of this disease, representing 18.4% of the total deaths by cancer [11]. Furthermore, the survival rate after five years of diagnosis for lung cancer is only 12% worldwide [24], which means that, in practice, being diagnosed today with lung cancer

amounts to a death penalty for patients. The reason for this staggering mortality rate is the near-absolute absence of apparent symptoms in patients of early lung cancer [10]. Typically, symptoms become evident only when the disease is highly advanced and other organs are already compromised. Consequently, the vast majority of lung cancers worldwide are diagnosed in stages III and IV, when the efficacy of existing treatments and hence the chances of survival are seriously compromised.

To improve the prognosis in patients with lung cancer, it is necessary to make the proper early diagnosis, which in turn requires an accurate understanding of medical images because of the visual evidence of this pathology. Nowadays, physicians use Computed Tomography (CT) scans to visualize the lungs of a patient and look for any life-threatening abnormality. However, because a pulmonary nodule is a rounded or irregular opacity that measures $\leq 30mm$ in diameter [14], finding them in a 3D medical image is a challenging task. Nodule detection by visual inspection is highly prone to error, even for specialists, resulting in the loss of between 43% and 52% of the nodules when evaluating the diagnostic images [18]. Once the nodules have been detected, the next step is the malignancy prediction of the findings. Specialists perform this task by visually inspecting the lesions and classifying them according to standardized descriptions of morphological characteristics, hence relying heavily on their own experience. Discrepancies among specialists are particularly concerning for this disease [6], as a single undetected or incorrectly classified malign nodule can compromise the life of the patient. This situation has spurred the appearance of machine learning techniques to assist specialists in early lung cancer detection. Accurate automated methods to perform this task would reduce the variations assessments made by different experts, providing a more robust measure of lung nodule presence, which is critical for early cancer diagnosis and treatment planning [12,22].

Progress in automated lung nodule detection and lung cancer diagnosis in the last decade is the result of a collective effort by a growing research community, which has undertaken the task of collecting and releasing large annotated datasets to train and evaluate quantitatively automated systems. A first pioneering effort was the LIDC/IDRI Database [6], which provided combined annotations by four experts of detected nodules for more than 1000 patients. Subsequently, the same data was used for the LUng Nodule Analysis (LUNA) Challenge in 2016 [16]. Concerning nodule classification, the first public challenge was the LUNGx SPIE-AAPM-NCI Lung Nodule Classification Challenge [5] in 2015, which provided malign/benign annotations for 60 nodules. In 2017, the Lung Nodule Malignancy Challenge [17] provided 1384 cases for nodule classification. In 2018, the ISBI Lung Nodule Malignancy Prediction Challenge [9] provided sequential low-dose CT (LDCT) scans at two screening intervals from the National Lung Screening Trial (NLST), with matched identified nodules from the same subject, for 100 patients. Finally, the 2020 Grand Challenge on Automatic Lung Cancer Management (LNDB) [19] focused on automatic clas-

sification of chest CT scans according to the 2017 Fleischner society pulmonary nodule guidelines for patient follow-up recommendation on 294 cases.

The availability of large-scale annotated datasets has spurred the development of deep learning techniques for nodule detection and lung cancer prediction. The current state-of-the-art is held by Ardila *et al.* [4], who obtained an overall 94.4% AUC on a National Lung Cancer Screening Trial test set with over 6000 patients. On the other hand, the best result obtained in the ISBI 2018 Lung Nodule Malignancy Prediction Challenge [9] was by [20], which obtains an AUC of 91.3% on the test set. Besides, [2] present a framework to detect lung nodules in four stages. Each stage allowed the refinement of the region to find a positive candidate and classify it as a nodule. Some other methods belong to the semi-supervised approach providing an interactive solution to the physician [7].

Although deep learning methods have pushed forward automated early lung cancer diagnosis in recent years, this task is still far from being solved and large-scale low-dose screening of the population is still years away from deployment. One of the main limitations for realistic lung cancer prediction of existing experimental frameworks is the formulation of the task itself, as all existing datasets and challenges seek to diagnose the disease *using exclusively visual data*. State-of-the-art approaches for lung cancer classification take as input a chest CT and produce a probability of cancer. This setup is radically different from clinical practice, which aggregates naturally multimodal information. Even though specialists evaluate by visual inspection standardized morphological characteristics of nodules for their classification, they also take into consideration all their knowledge of the context and the patient's history. As an example, a radiologist will not study in the same way an image from a healthy child and one from a person with a 20 pack-year smoking history.

In this paper, we present the **LU**ng **CA**ncer **S**creening (LUCAS) Dataset, the first multimodal experimental framework for early lung cancer diagnosis. We collected a large dataset with low-dose chest CT scans of 830 patients in a real-world screening scenario in which only a small fraction of the cases is diagnosed with lung cancer, and the rest, though sane, belong to a population that is exposed to risk factors. We complement this visual information with anonymized clinical data and additional information from the radiologists' reports. The goal of the LUCAS dataset is to serve as a testbed to assess the relative importance and complementarity of the different modalities of data for lung cancer diagnosis.

In order to assess the difficulty of the task in the LUCAS Dataset and to set a baseline for future reference, we develop a deep learning technique that combines visual and clinical data for lung cancer prediction. Given the highly unbalanced nature of the detection task we address, we model it as a detection problem and we evaluate our results with the point that maximizes the Precision-Recall curve (F-score) [13,15], a standard metric in computer vision [3] that is more stringent than the AUC-ROC used by existing datasets.

Our results show that both modalities are complementary for an accurate diagnosis. Furthermore, benchmarking results in a realistic setup with the F-score reveals the true complexity of the task, as the performance of our multimodal

system on the test set is only 25%. This sobering result implies that automated lung cancer screening is still a challenging open problem. To promote the appearance of a new wave of multimodal automated methods for lung cancer diagnosis, we make publicly available all the resources of this project [3].

2 Lung Cancer Screening (LUCAS) Dataset

To create the LUCAS dataset, we partnered with a large healthcare institution that provides treatment to patients with a wide variety of diseases and with different risks of lung cancer. For the data acquisition process, we first collect all the chest CT scans that had been done in the healthcare institution during a period of one year, regardless of the diagnosis of the patients. Thus, our collection process ensures that the dataset is representative of a real clinical scenario, with patients having different risks of developing lung cancer, and that the data mimics the incidence of lung cancer in the population. We also collect the clinical report associated with each of the CT scans to have an integral understanding of the context regarding the medical history of the patient.

For each of the patients in the LUCAS dataset, we have the latest CT scan that was performed on the patient as well as the clinical report in which the physician explains the findings related to that CT scan and states if the patient has cancer or not. We anonymize all of the CT scans and clinical reports according to the established standards. We process this information to create a multimodal framework that includes visual information as well as a set of relevant biomarkers that might indicate risk factors.

2.1 Visual information

The LUCAS dataset contains 830 low-dose chest CT scans from patients in a real-life setting. 72 of these patients are diagnosed with cancer by an expert physician. Nonetheless, most of these patients have a respiratory disease or are at high risk of developing lung cancer. The diversity in patient's diagnosis makes identifying visual patterns a complex task. We select 20% of the patients for the testing set and ensure that both groups share the same proportion of patients with and without cancer.

2.2 Biomarkers

For each patient in the LUCAS dataset, we have a clinical report associated with the patient's last CT scan. However, the information in these reports varies according to the level of detail registered by the physician. For this reason, we turn the clinical reports into sets of structured information to facilitate automated interpretation. We translate the reports into biomarkers that include relevant information for lung cancer diagnosis. Aided by expert physicians, we select characteristics that are transversal to every clinical report.

[3] https://github.com/BCV-Uniandes/LUCAS

Table 1. Categories of Biomarkers in the LUCAS Dataset

Category	Biomarkers
Cancer Related Factors	Cancer History, Presence of Pulmonary Nodules, Pulmonary Nodule Characteristics, Presence of Pulmonary Masses, Characteristics of Pulmonary Masses.
Clinical History	Respiratory History, Previous CT scans, Adenomegaly, Thoracic Pain, Pleural Effusion.
Visual Analysis	Granulomas, Pulmonary Parenchymal Consolidation, Presence of Tree-in-bud, Opacities.

The first category includes factors that are directly correlated with lung cancer such as medical history of cancer or factors that are a direct consequence of lung cancer like presence of pulmonary nodules. The second category is biomarkers regarding the patient's clinical history. The last category corresponds to visual analysis biomarkers. This category is composed by visual aspects of the CT scan that were highlighted by the physician in the report. Table 1 shows a description of the biomarkers in each category.

2.3 Task

For this dataset, we propose to study the screening task in which the distribution of the data resembles the real-life class imbalance. In this case, we have 8% of positive samples in the entire dataset. Additionally, as the evaluation metric, we propose to study this problem with the maximal F-score on the Precision-Recall curve [13,15].

3 Baseline Approach

3.1 Image pre-processing

Variations in the voxel spacing of data may affect CNNs understanding of the images. For this reason, we resample the volumes to the median voxel spacing of the entire dataset using spline interpolation of third degree. In addition, the images are cropped along the depth dimension to include only the lungs of the patients. Finally, we perform a z-score normalization based on the statistics calculated for lung nodules in the Medical Segmentation Decathlon (MSD) [21] task for lung nodule segmentation.

3.2 Method

We train three methods to predict the probability of cancer for each patient. The models vary according to the modality of information used as input for the classification task. We test the effectiveness of using only visual information from the CT scans, using the biomarkers obtained from the clinical reports, and combining multimodal information.

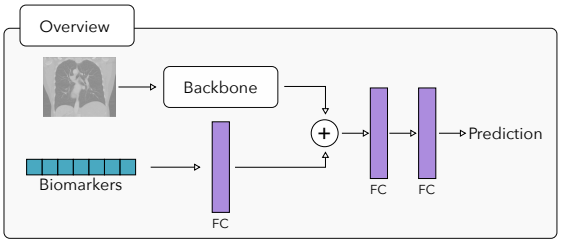

Fig. 1. Overview of the multimodal classification method. Our model extracts features from the diagnostic image and clinical report of a patient. Afterward, two Fully Connected (FC) layers combine the information to predict the probability of cancer.

Image-based approach: we use a classification network with a backbone pretrained for lung nodule segmentation using the lung dataset from the MSD, and two fully connected layers to obtain the probability of cancer. The backbone has five stages with depthwise separable convolutions to reduce the computational cost derived from processing 3D images, and a strided convolution to reduce the image size. The number of feature maps is set to 32 in the first stage and is doubled after every dimensionality reduction, making sure that the maximum number of feature maps is 512. Also, to alleviate the issues derived from using small batches, we use instance normalization instead of the standard batch normalization.

Biomarkers: in this setting we use a multilayer perceptron to learn the relation between biomarkers and the relative importance of each one of them. The model contains a fully connected layer to encode the inputs followed by the same classification layers used for image classification.

Multimodal approach: for the final model, we integrate both approaches for the individual modalities to obtain visual features from the CT scans and relevant information from the clinical reports. We perform the encoding stage for images and biomarkers in parallel and concatenate the resulting features to learn a joint representation using fully connected layers. Finally, the patient is classified according to the predicted probability of having cancer. An overview of the proposed pipeline is shown in Figure 1

3.3 Training details

We train our model for 40 epochs using Adam optimizer with weight decay of $1e-5$ and an initial learning rate of $1e-3$. The learning rate is reduced by a factor of 0.1 if the validation loss has not decreased in the previous 10 epochs.

Taking into account the large imbalance in the dataset, during training we assign higher probabilities of being selected to patients with cancer. By doing so, only a fraction of the negative samples is randomly selected every epoch, resulting in a natural data augmentation strategy.

Table 2. Results of the three variants of our baseline algorithm on the LUCAS Dataset.

Metric	Images	Reports	Multimodal Data
ROC	0.513	0.674	**0.712**
F score	0.207	0.162	**0.250**

4 Results

In order to assess quantitatively the difficulty of the lung cancer detection task on the LUCAS Dataset, as well as the relative importance of the different information modalities, we evaluate the three variants of our baseline algorithm on the test set.

Table 2 presents the results for both the ROC-AUC, the metric used in previous lung cancer datasets, and the maximal F-score on the Precision-Recall curve, the performance measure we adopt in this paper. The difference in absolute scores for the two metrics highlights the more stringent nature of the F-score and its appropriateness for detection tasks, as, in contrast to the ROC-AUC, it does not take into account true negatives in the computation. However, in both metrics, our multimodal baseline clearly outperforms the two versions of the system with only one modality. This result indicates that the two modalities provide complementary information and that our method is capable of taking benefit for improved detection. However, a closer look at the scores reveals that the maximal F-score is only 25% for the combined system, suggesting that lung cancer detection in the realistic setting of the LUCAS Dataset is still a very challenging problem, even in the presence of multimodal biomarkers.

In order to gain further insights on the results, we make use of the Toolkit for analyzing and visualizing challenge results [23] an evaluation framework that was designed to measure statistical significance of performance among different algorithms on biomedical machine learning challenges, and that was used to analyze the EndoVis 2019 Challenge results. Since the toolbox was created for a setting in which a performance metric such as the Dice Index is used to score softly algorithmic results rather than with binary detection labels, we use the detection probability as a score for positive instances and its complement for negative instances.

The main plots from the significance analysis are reported in Figs. 2 and 3. Figure 2 shows different possible rankings for the three algorithms, all based on the same individual scores. We can observe that the multimodal baseline is consistently ranked first, while the versions with a single modality can switch places depending on the specific ranking mechanism. This result underscores again the complementary nature of visual and clinical data, as well as the appropriateness of our multimodal system for leveraging it. It is also consistent with the apparent discrepancy in the ranking of individual modalities with the two metrics in Table 2. Figure 3 presents the dot-and-box plots for the individual scores, revealing a much tighter distribution for the multimodal system, and hence providing supporting evidence for the statistical significance of our results.

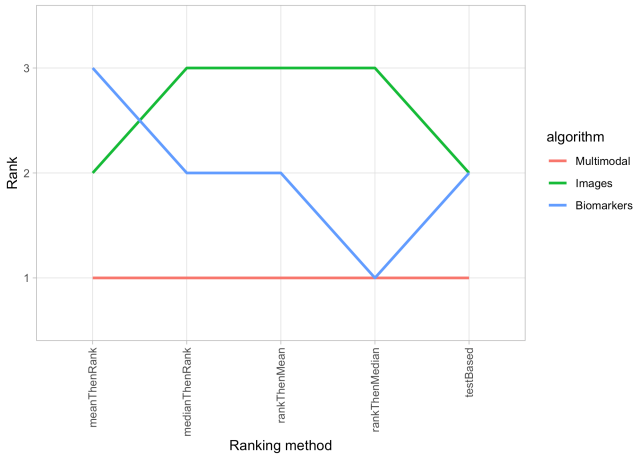

Fig. 2. Ranking robustness on three variants of our algorithm for the LUCAS Dataset. Using multimodal information proves to be more robust under every ranking.

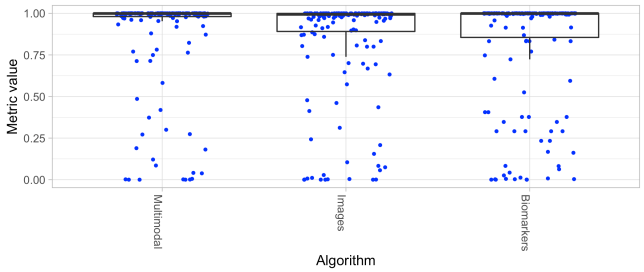

Fig. 3. Dot-and-box plot for the three variants of our algorithm on the LUCAS Dataset.

5 Conclusions

We present the LUCAS Dataset, the first experimental testbed for lung cancer detection with multimodal biomarkers. In addition to low-dose CT scans of hundreds of patients in a realistic clinical screening setup, we provide key clinical data from anonymized records and radiology reports to enrich the analysis. We develop a multimodal deep neural network as a strong baseline for future reference, and we show empirically the complementary nature of visual and clinical data for lung cancer detection. We hope that the availability of our experimental framework will enable the development of new generations of multimodal techniques and the exploration of new ideas for early lung cancer diagnosis.

Acknowledgments: This project was partially funded by the Google Latin America Research Awards (LARA) 2019.

References

1. Abbosh, C., Birkbak, N.J., Wilson, G.A., Jamal-Hanjani, M., Constantin, T., Salari, R., Le Quesne, J., Moore, D.A., Veeriah, S., Rosenthal, R., et al.: Phylogenetic ctdna analysis depicts early-stage lung cancer evolution. Nature **545**(7655), 446–451 (2017)
2. Alilou, M., Kovalev, V., Snezhko, E., Taimouri, V.: A comprehensive framework for automatic detection of pulmonary nodules in lung ct images. Image Analysis & Stereology **33**, 13 (03 2014). https://doi.org/10.5566/ias.v33.p13-27
3. Arbelaez, P., Maire, M., Fowlkes, C., Malik, J.: Contour detection and hierarchical image segmentation. IEEE transactions on pattern analysis and machine intelligence **33**(5), 898–916 (2010)
4. Ardila, D., Kiraly, A.P., Bharadwaj, S., Choi, B., Reicher, J.J., Peng, L., Tse, D., Etemadi, M., Ye, W., Corrado, G., Naidich, D.P., Shetty, S.: End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nature Medicine **25**(6), 954–961 (May 2019). https://doi.org/10.1038/s41591-019-0447-x
5. Armato, S.G., Drukker, K., Li, F., Hadjiiski, L., Tourassi, G.D., Engelmann, R.M., Giger, M.L., Redmond, G., Farahani, K., Kirby, J.S., Clarke, L.P.: LUNGx challenge for computerized lung nodule classification. Journal of Medical Imaging **3**(4), 044506 (Dec 2016). https://doi.org/10.1117/1.jmi.3.4.044506
6. Armato III, S.G., McLennan, G., Bidaut, L., McNitt-Gray, M.F., Meyer, C.R., Reeves, A.P., Zhao, B., Aberle, D.R., Henschke, C.I., Hoffman, E.A., et al.: The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. Medical physics **38**(2), 915–931 (2011)
7. Astaraki, M., Toma-Dasu, I., Smedby, Ö., Wang, C.: Normal appearance autoencoder for lung cancer detection and segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 249–256. Springer (2019)
8. AstraZeneca, I.: Summary of product characteristics (2007)
9. Balagurunathan, Y., Farahani, K., Goldgof, D., Hadjiyski, L., Kalpathy-Cramer, J., McNitt-Gray, M., Napel, S.: Isbi 2018 - lung nodule malignancy prediction challenge. http://isbichallenges.cloudapp.net/competitions/15 (2018)
10. Beckles, M.A., Spiro, S.G., Colice, G.L., Rudd, R.M.: Initial evaluation of the patient with lung cancer: symptoms, signs, laboratory tests, and paraneoplastic syndromes. Chest **123**(1), 97S–104S (2003)
11. Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., Jemal, A.: Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians **68**(6), 394–424 (Sep 2018). https://doi.org/10.3322/caac.21492
12. del Ciello, A., Franchi, P., Contegiacomo, A., Cicchetti, G., Bonomo, L., Larici, A.R.: Missed lung cancer: when, where, and why? Diagnostic and Interventional Radiology **23**(2), 118–126 (Mar 2017). https://doi.org/10.5152/dir.2016.16187
13. Flach, P., Kull, M.: Precision-recall-gain curves: Pr analysis done right. In: Advances in neural information processing systems. pp. 838–846 (2015)
14. Hansell, D.M., Bankier, A.A., MacMahon, H., McLoud, T.C., Müller, N.L., Remy, J.: Fleischner society: Glossary of terms for thoracic imaging. Radiology **246**(3), 697–722 (Mar 2008). https://doi.org/10.1148/radiol.2462070712

15. Hariharan, B., Arbeláez, P., Bourdev, L., Maji, S., Malik, J.: Semantic contours from inverse detectors. In: 2011 International Conference on Computer Vision. pp. 991–998. IEEE (2011)
16. Jacobs, C., Setio, A.A.A., Traverso, A., van Ginneken, B.: Luna16. https://luna16.grand-challenge.org (2016)
17. Kaggle: Data science bowl 2017. https://www.kaggle.com/c/data-science-bowl-2017 (2017)
18. Memon, W., Haider, Z., Memon, W., Idris, M., Kashif, N., Idris, S.M., Sajjad, Z., Saeed: Can computer assisted diagnosis (CAD) be used as a screening tool in the detection of pulmonary nodules when using 64-slice multidetector computed tomography? International Journal of General Medicine p. 815 (Dec 2011). https://doi.org/10.2147/ijgm.s26127
19. Pedrosa, J., Ferreira, C., Aresta, G.: Grand challenge on automatic lung cancer patient manager. https://lndb.grand-challenge.org/ (2020)
20. Pérez, G., Arbeláez, P.: Lung cancer prediction (2018), https://biomedicalcomputervision.uniandes.edu.co/index.php/research?id=33
21. Simpson, A.L., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., van Ginneken, B., Kopp-Schneider, A., Landman, B.A., Litjens, G.J.S., Menze, B.H., Ronneberger, O., Summers, R.M., Bilic, P., Christ, P.F., Do, R.K.G., Gollub, M., Golia-Pernicka, J., Heckers, S., Jarnagin, W.R., McHugo, M., Napel, S., Vorontsov, E., Maier-Hein, L., Cardoso, M.J.: A large annotated medical image dataset for the development and evaluation of segmentation algorithms. CoRR **abs/1902.09063** (2019), http://arxiv.org/abs/1902.09063
22. Way, T., Chan, H.P., Hadjiiski, L., Sahiner, B., Chughtai, A., Song, T.K., Poopat, C., Stojanovska, J., Frank, L., Attili, A., Bogot, N., Cascade, P.N., Kazerooni, E.A.: Computer-aided diagnosis of lung nodules on CT scans:. Academic Radiology **17**(3), 323–332 (Mar 2010). https://doi.org/10.1016/j.acra.2009.10.016
23. Wiesenfarth, M., Reinke, A., A.L., L., Cardoso, M., Maier-Hein, L., Kopp-Schneider, A.: Methods and open-source toolkit for analyzing and visualizing challenge results. arXiv preprint arXiv:1910.05121 (2019)
24. Wong, M.C., Lao, X.Q., Ho, K.F., Goggins, W.B., Shelly, L.: Incidence and mortality of lung cancer: global trends and association with socioeconomic status. Scientific reports **7**(1), 1–9 (2017)