Eurographics Workshop on Visual Computing for Biology and Medicine (2022)
T. Schultz and H.-Y. Wu (Editors)

Short Paper

Predicting, Analyzing and Communicating Outcomes of COVID-19 Hospitalizations with Medical Images and Clinical Data

Oliver Stritzel and Renata Georgia Raidou

TU Wien, Austria

Abstract

*We propose PACO, a visual analytics framework to support the prediction, analysis, and communication of COVID-19 hospitalization outcomes. Although several real-world data sets about COVID-19 are openly available, most of the current research focuses on the detection of the disease. Until now, no previous work exists on combining insights from medical image data with knowledge extracted from clinical data, predicting the likelihood of an intensive care unit (ICU) visit, ventilation, or decease. Moreover, available literature has not yet focused on communicating such results to the broader society. To support the prediction, analysis and communication of the outcomes of COVID-19 hospitalizations on the basis of a publicly available data set comprising both electronic health data and medical image data [SSP*21], we conduct the following three steps: (1) automated segmentation of the available X-ray images and processing of clinical data, (2) development of a model for the prediction of disease outcomes and a comparison to state-of-the-art prediction scores for both data sources, i.e., medical images and clinical data, and (3) the communication of outcomes to two different groups (i.e., clinical experts and the general population) through interactive dashboards. Preliminary results indicate that the prediction, analysis and communication of hospitalization outcomes is a significant topic in the context of COVID-19 prevention.*

CCS Concepts

• ***Human-centered computing*** → ***Visual Analytics;*** • ***Applied computing*** → ***Life and medical sciences;***

1. Introduction

COVID-19 is a respiratory disease that turned into a pandemic in 2020. While the main danger of the disease is the actual illness, it has also fueled social disruption through fake news and alternative facts—hindering strategies to combat the pandemic's spread. This could have been mitigated through a more offensive information strategy in highly trusted media, and it seems increasingly important to communicate novel COVID-19-related insights gained by scientific institutions to the general population in an understandable manner [FDEO20].

Several real-world data sets about COVID-19 are openly available. With the disease affecting the lungs, data sets often comprise solely chest X-rays or inherit other medical images, such as computed tomography (CT) scans, and/or clinical data from electronic health records. This vast information can be used to train models for disease detection [ZCH*20, LHL*20] or for the prediction of high-risk patients [ZCH*20]. Until now, no work exists on combining insights from medical images with knowledge extracted from clinical data for COVID-19, predicting the likelihood of an intensive care unit (ICU) visit, ventilation, or decease. Moreover, available literature has not focused on communicating such results to the broader society—especially, laypeople.

Although the prediction of the status of a patient infected by COVID-19 could be simply evaluated quantitatively using retrospective data, the communication of the prediction outcomes requires an additional visual presentation of the insights to the target users. Visual analytics offers a welcoming opportunity for providing significant support through visuals that communicate information—helping clinical experts to save time and resources, while also facilitating the understanding of prediction models, and providing risk communication tools for the general population.

The *contribution* of this work is a visual analytics framework that supports the prediction, analysis and communication of the outcomes of COVID-19 hospitalizations on the basis of a publicly available data set [SSP*21]. The main components of our framework include the analysis and prediction of hospitalization outcomes using electronic health data and medical image data, and the communication of the outcome prediction for two different user groups (namely, clinical experts and the general population).

2. Related Work

Applications for the communication of the status of patients have been proposed already numerous times, and is a very active field of research. Examples include previous work providing decision making support and patient cohort exploration. Recent work by Furmanová et al. addresses the exploration, analysis, and prediction of pelvic organ variability to support decision making with regard to tumor treatment [FMCM*21]. Floricel et al. went one step further, developing an environment for visual analysis and knowledge discovery for longitudinal cancer therapy symptom data [FNB*21]. In

DOI: 10.2312/vcbm.20221196

a different application, Bernold et al. proposed a dashboard targeting predictive analytics on the basis of in-patient rehabilitation data from a large cohort of 46,000 patients [BMGR19]. Through the pandemic, a lot of research has been shifted towards developing applications with communication purposes in mind, focusing on COVID-19 detection using medical images [ZR21] or electronic health data records [ZCH*20]. Furthermore, medical biomarkers have been defined in multiple publications regarding COVID-19 disease progression [PMR*20] along the typical biomarkers analysis in cancer research [ZVA*20]. To the best of our knowledge, there is no previous work that combines multiple COVID-19 hospitalization data sources, while communicating insights through interactive dashboards to different target groups.

3. Data – Users – Tasks Analysis

■ **Data:** We employ the Stony Brook University COVID-19 Positive Cases Data set [SSP*21], which includes a variety of medical information in form of tabular data at a per-patient level (in total 131 features for electronic health records, including information on demographics, pre-existing conditions, and results from medical tests during the hospital stay, such as blood oxygen tests). Patients included in the data set ($N = 1384$) were all hospitalized and tested positive using a polymerase chain reaction (PCR) test for COVID-19. The data set contains missing values, which need to be dealt with, e.g., with imputation methods. Additionally, the data set contains chest X-ray images without segmentation masks of the lungs or other structures. Multiple images may be available for a patient and are often available per day, and in most cases, at least two different contrast settings have been used. After cleaning up the data, a final number of $N = 1279$ patients and 4728 X-ray images is included. Among these patients, 174 (13.6%) are deceased, 257 (20.1%) have been admitted to ICU, and 213 (16.7%) were ventilated during their hospitalization stay.

■ **Users:** We separate the general population into two categories:

(U1) Medical experts and clinicians, who are interested in decision making support for the treatment of upcoming hospitalized patients. This includes comparing and filtering for similar patients based on electronic records and preconditions, which could also be used for prediction or risk perception dialogues in individual patient treatments.

(U2) General population, i.e., lay users without specific backgrounds in medicine or data analytics. Insights can be communicated to this group, with the purpose of increasing risk perception and better support of health-related measures. Here, the biggest challenge is the varying level of visualization literacy of the users.

■ **Tasks:** We support the users in accomplishing these three tasks:

(T1) Automated segmentation requires the segmentation of X-ray images, feature extraction thereof, and correlation with radiomic features to support the identification of potential biomarkers. This task requires also the **processing of the clinical data**.

(T2) Prediction of disease outcomes requires the prediction of disease outcomes and a comparison to state-of-the-art prediction scores for both data sources, i.e., medical images and clinical data.

(T3) Outcomes communication indicates that the prediction of disease outcomes needs to be communicated to the aforementioned users through an interactive interface. Each group focuses on different aspects of the information space to discover new knowledge.

4. Predicting, Analyzing and Communicating Outcomes of COVID-19 Hospitalizations with PACO

We designed and developed PACO, a visual analytics dashboard to support the Prediction, Analysis and Communication of COVID-19 hospitalization Outcomes. The workflow of PACO is depicted in Figure 1. PACO is developed in Python, using PyTorch, scikit-learn and imbalanced-learn. For the medical images, we used pydicom, scikit-image and PyRadiomics. For the front-end development, Streamlit was used together with Plotly. The implementation is made available on GitHub.

■ **Automated Segmentation and Clinical Data Processing:**

Lung segmentation from X-rays: To train a lung segmentation network from publicly available masked X-ray images, we focus on investigating previously proposed architectures [IZ18, IS18]. Islam et al. [IZ18] make use of openly available data sets (Montgomery County and Shenzen Lung data) to train a Unet network specialized for lung segmentation. *TernausNet* (*VGG11 Unet*) has been widely used for lung segmentation [IS18] as an improvement to the traditional *Unet*, by generating better features and boosting its performance. It can be initialized with a pre-trained network in a warm-start scenario, and the authors showed promising results when using the model pre-trained on ImageNet.

In all previously investigated approaches, the medical images differ from the ones in our data set, as ours are less contrasted and with heavy artifacts caused by cables or the positioning of the patient. To make our data set more compatible with the training data, preprocessing is applied, in the form of a Gaussian blur filter followed by an adaptive histogram equalization. Subsequently, we train five models shown in Table 1. Given that our data set is less clean than the training, we add additional rotational data augmentation during the training phase to make the model more robust. In addition, random cropping, zooming, and shifting are applied to the data during training to increase variance. We also opt for testing both *AdamW* optimizer and *Adam*, as the former tends to yield better training loss and generalizes much better than the latter.

To form a well-founded decision about the model despite the lack of ground truth, a quantitative assessment is designed based on previous work on reverse classification accuracy by Valindria et al. [VLB*17]. While the reverse accuracy is based on a model trained on only one sample that is presented to an online-segmentation system, the data available in this work allows for the creation of many reference segmentations to train a new model. Using parts of the original training data sets (Montgomery County and Shenzen Lung data) with ground truth available as a validation set, the usage of standard metrics like Dice and Jaccard is enabled again. Figure 2 schematically depicts our evaluation process, which results in Model A from Table 1 (*TernausNet with Adam Optimizer*) being the preferred model for the segmentation, with a loss of 0.0905±0.0560, Jaccard Index of 0.8858 ± 0.0570, and Dice Coefficient of 0.9384 ± 0.0342.

Radiomic feature extraction: In the next step, PyRadiomics is

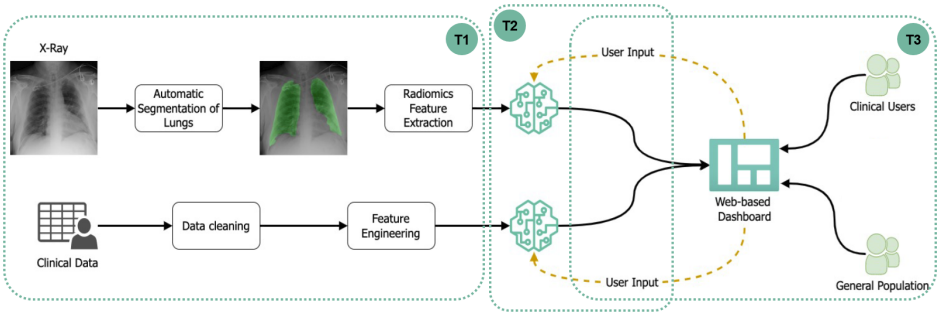


Figure 1: *The workflow of* PACO *for predicting, analyzing and communicating outcomes of COVID-19 hospitalizations with medical images and clinical data, including a link to the main tasks:* ***(T1)*** *Automated segmentation,* ***(T2)*** *Prediction and* ***(T3)*** *Outcomes communication.*

Name	Design	Rotation	Optmizer	Early stop
Model A	TernausNet	None	Adam	None
Model B	TernausNet	± 30	AdamW	10%
Model C	TernausNet	± 30	Adam	10%
Model D	TernausNet	± 25	Adam	20%
Model E	UNet	± 25	Adam	20%

Table 1: *Deep learning models [IS18, IZ18] trained on Montgomery County and Shenzen Lung data.*

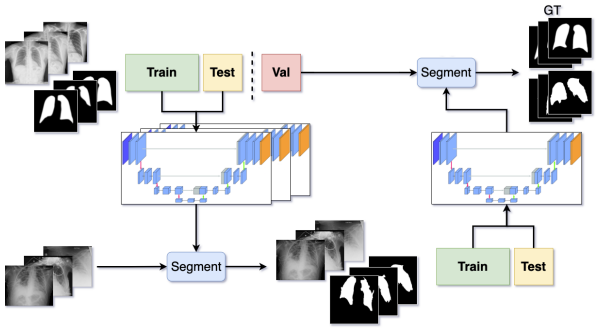


Figure 2: *Model evaluation strategy without ground truth, inspired by Valindria et al. [VLB*17].*

used to extract radiomic features from the chest X-rays and their respective segmentation masks. This can provide important information regarding potential biomarkers [PMR*20]. First, the lung masks are cleaned, removing additional wrongly segmented areas, based on the two biggest contours as generated by contour detection on binarized gray scale masks. Then, the remaining area is separated into left and right lung using positional information. Radiomic feature extraction with stacked features for all feature classes leads to a total of 204 features, with 102 for each lung.

Clinical data cleaning: To prepare the clinical data, we initially transform them to a desired form. For example, we *transform* the variable that represents the outcome of the hospitalization from a string format to a numerical, leading to eight target variables that define our multi-label classification setting. Additionally, duplicate and redundant variables are *removed* (e.g., Body Mass Index > 35 boolean fields were removed, if also present as numeric values). Then, *one-hot encoding* is applied to some features, where we convert each categorical value into a numerical one and assign a binary value of 0 or 1 to it. For example, smoking status would be 0 if the patient does not smoke and 1, otherwise. Also, targets and data columns that could implicitly indicate the outcome of any of the target variables (e.g., number of ventilation days imply that a patient was ventilated) are removed. Finally, we *scale* the remaining numerical variables by removing the mean and scaling to unit variance. To address data missingness, we have applied and compared *different imputation strategies* (mean, median, regression, k-nearest neighbor imputation) [VB18]. After a thorough quantitative assessment using a leave-one-out method on complete data, where we computed the Silhouette Score and the Calinski Harabasz (CH) criterion, and an additional qualitative comparison of the resulting distribution plots, kNN imputation with $k = 5$ is chosen as the best fit. After imputation, the clinical data together with the radiomic features are *scaled* to account for changes in the data distribution. This results into a 284-dimensional feature vector, where the 204 are the radiomic features of the previous section and the remaining 80 are clinical features.

Clinical feature engineering: Generating predictions requires as the first step to *cluster* patient records into semantically reasonable groups depending on their hospitalization outcomes and their general clinical state, as denoted by the available clinical data. Each patient is described with 284 features, i.e., 204 radiomic and 80 clinical. To work with this high-dimensional space, prior to clustering, we need to apply *dimensionality reduction*, namely t-SNE [VdMH08]. For the clustering, we applied and compared the effect of different *clustering* methods (k-means, DBSCAN and Ward hierarchical clustering) [RM05] using the Silhouette Score and the CH criterion against the ground truth outcomes (i.e., deceased, ICU admission, ventilated, and hospitalized). A quantitative evaluation of the different alternatives suggests that the most robust option is yielded when employing k-means with $k = 4$, returning 4 clusters.

After analyzing the characteristics of the four resulting clusters, we summarize them as follows: *Cluster 1* comprises *healthy young to middle-aged patients*, unlikely to have preconditions, with the lowest decease rate and the shortest hospital stay on average, low ventilation rate, and mainly women. *Cluster 2* contains *less healthy young to middle-aged patients*, unlikely to have preconditions, with large hospital stays, high chance of ICU admission and ventilation, and mainly men. *Cluster 3* consists of *elderly people with high risk*,

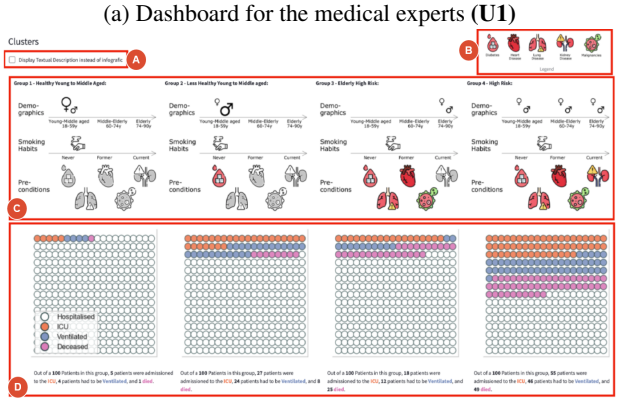
likely to have preconditions such as diabetes, heart issues or malignancies, with the lowest BMI and the lowest rate of never-smokers, and the highest (and fastest) decease rate. *Cluster 4* includes *high-risk patients with pre-existing conditions* from all age groups, who have a long and severe hospital stay and were mostly admitted to ICU, ventilated, and deceased.

■ **Prediction of Disease Outcomes:** The data set inherits multiple variables that could be chosen as prediction targets. In this work, we choose the outcome of the disease as: deceased, dismissed from the hospital, ventilated during the hospital stay, and admitted to the ICU. These can be defined as separate binary classification cases, which would lead to several representation issues for the dashboard and the users, as separate classifiers would lead to independent predictions. To avoid this, we combine the classes into a multi-class classification setting. This reduces complexity and enables communication of results in form of natural probabilities for an outcome to the potential users for the dashboard. Yet, this introduces several other problems like class imbalances, for which different strategies are adopted—namely, using balanced weights where applicable, random oversampling, random undersampling, and Synthetic Minority Oversampling Technique (SMOTE) [BSGR03].

The data are split into 70% train and 30% validation and are not imputed prior to splitting to avoid information leakage. Training data are used within 8-fold cross-validation, as stratified folding is limited to the lowest number of records for minority classes, which was 8. For each fold, the clean data are used to train five classifiers: Random Forest ($n = 10$, $max\_depth = 5$), Logistic Regression ($max\_iter = 250$), Support Vector Machine (SVM), XGBoost, and Multi-layer Perceptron (MLP, $max\_iter = 500$, $lr = adaptive$). Subsequently, we assess which approach yields the most satisfactory results. We reuse metrics proposed by other works to be able to compare our results directly [ZCH*20, LHL*20]. In addition to accuracy, recall and ROC–AUC (one vs. one), we quantify balanced accuracy. Accuracy is a non-optimal metric for imbalanced data sets, not taking class weights into account and thus tending to provide too optimistic results for classifiers that are biased towards the majority class. Balanced accuracy takes this into account. The best classifier overall is SVM with balanced class weights, as it has the highest ROC–AUC tied with XGBoost. SVM is superior though in balanced-accuracy and has a lower standard deviation over the training folds. SVM is trained again with k-fold cross-validation to tune for optimal hyper-parameters. The best ROC–AUC was found for $C = 25$ with a linear kernel and seed 0, with a ROC of 0.795 ± 0.063 on the training set and 0.78 on the validation set.

■ **Outcomes Communication:** Medical experts need to focus on the prediction of the disease progression based on the available data, while the general population would care for "what–if" scenarios applied to personal or familial data. For example, the former would look into potential outcomes of a new incoming patient, while the latter would use it to find out how their general health status might influence (or not) the likelihood of hospitalization. The medical tasks require overview strategies into the multivariate electronic health data and medical images, while for the general population, the data should not be presented in its raw form to avoid overwhelming views. Interaction possibilities are needed by all target groups to enable selection and filtering.

(a) Dashboard for the medical experts **(U1)**



(b) Dashboard for the general population **(U2)**

Figure 3: *Dashboards of* PACO *for the two user groups.*

For medical experts and clinicians (U1): The entire segmentation and prediction part of the pipeline is conducted in the background, and only the results can be accessed by the medical expert or clinician. The user starts by deciding the type of patients to include in the analysis (Figure 3 (a,A)). Then, filtering and querying can be performed (Figure 3 (a,B)), where specific patient characteristics, such as gender, can be included. Subsequently, we provide an overview of the entire cohort on a scatterplot that represents the days of hospitalization vs. the number of X-rays acquired per patient. An additional colorcoding of the data points represents the eight hospitalization outcome classes (Figure 3 (a,C)). Here, the clustering outcomes (see four clusters above) are not communicated to the user, and the eight types of patients are determined from the past cohort (e.g., deceased, or ICU, or ICU+ventilated). To obtain more details on demand, patients can be selected in the scatterplot (Figure 3 (a,C)) or manually (Figure 3 (a,D)), and their medical images will be shown (Figure 3 (a,E)). Specific timepoints of the hospitalization can be filtered (Figure 3 (a,F)), while the predicted segmentations can be overimposed on the X-rays ((Figure 3 (a,G)). The user can further load images of a new patient to predict the hospitalization outcome, based on similar past patients.

For the general population (U2): The dashboard for the general population is simplified, as—to support all visualization literacy levels—we provide all information textually or using infographics. The users can select which of the two ways they prefer (Figure

3 (b,A)). If the users choose infographics, a legend with schematic depictions of the potential underlying conditions is shown to them (Figure 3 (b,B)). Subsequently, the four groups of patients, as resulting from the clustering step, are also communicated (Figure 3 (b,C)). Here, we prefer to use the clustering outcome, as these groups are more descriptive than the raw data classes. Additional information for each group is visually abstracted and communicated to the user, including demographics, smoking habits, and preconditions. Here, glyphs are employed (e.g., for genders denoted as ♂ and ♀), while for the preconditions we use Focus+Context, to indicate which are encountered (e.g., group 1 has no precondition, as all are greyed out, while group 2 has diabetes). Finally, we use a scatterplot-like representation, where data points are placed on a grid and color coded depending on whether they were hospitalized, deceased, ventilated, or in the ICU (Figure 3 (b,D)). This is accompanied by an additional textual description at the bottom of the dashboard. In all cases, juxtaposition of the groups is preferred to show all cases comparatively. The users can further input personal information (e.g., age, smoking status, preconditions) to predict a personalized hospitalization outcome. Using medical images, if available, as the input is also supported. Here, the user will additionally obtain a summary of top five most similar patients.

5. Evaluation

We conducted an initial user study with members from the general population ($N = 6$, 3 male, mean age 29), and two fictional patient cases were prepared to validate the dashboard design for user group **U2**. For example, one case was: *Till is 28 years old and has diabetes. He is a current smoker and not very sporty. In which risk group would you assign him? What would be his outcome prediction if he was hospitalized for COVID-19?* Before conducting the cases, the users had time to get familiar with PACO and to ask questions. All participants were able to solve the two cases correctly and to assign the patients to the correct groups. Using the patient's—or even personal—data as input for the prediction was also sufficiently conducted and interpreted. One of the two cases included a patient precondition (i.e., coronary artery disease), which was misinterpreted by some participants and led them to not entirely accurate results. All interviewees positively outlined the possibilities of using their own data and interacting with the dashboard. Four mentioned that the infographic was aesthetically pleasing, intuitive, and easy to understand. The other two found it confusing or commented that it required more attention. As a solution, we also provided textual feedback, but this was only used by one participant. Three interviewees wished for a more granular differentiation between the age groups. Two interviewees would have found it helpful to have information about the patients' fitness, and one suggested including vaccination information to educate skepticals.

6. Conclusions and Future Work

In this work, we proposed PACO, a dashboard to support the prediction, analysis and communication of COVID-19 hospitalization outcomes to two user groups. In our future work we would like to integrate vaccination data in the prediction model, which could be further reworked to predict other variables, such as the length of stay. Finally, we would like to extend the evaluation to both groups to obtain insights into PACO's practical usefulness.

References

[BMGR19] Bernold G., Matkovic K., Gröller M. E., Raidou R.: preha: Establishing precision rehabilitation with visual analytics. In *Eurographics Workshop on Visual Computing for Biology and Medicine (2019)* (2019), pp. 79–89. 2

[BSGR03] Barandela R., Sánchez J. S., Garcia V., Rangel E.: Strategies for learning in class imbalance problems. *Pattern Recognition 36*, 3 (2003), 849–851. 4

[FDEO20] French J., Deshpande S., Evans W., Obregon R.: Key Guidelines in Developing a Pre-Emptive COVID-19 Vaccination Uptake Promotion Strategy. *International Journal of Environmental Research and Public Health 17*, 16 (2020). 1

[FMCM*21] Furmanová K., Muren L. P., Casares-Magaz O., Moiseenko V., Einck J. P., Pilskog S., Raidou R. G.: PREVIS: Predictive visual analytics of anatomical variability for radiotherapy decision support. *Computers & Graphics 97* (2021), 126–138. 1

[FNB*21] Floricel C., Nipu N., Biggs M., Wentzel A., Canahuate G., Dijk L. V., Mohamed A., Fuller C., Marai G. E.: THALIS: Human-Machine Analysis of Longitudinal Symptoms in Cancer Therapy. *IEEE Transactions on Visualization and Computer Graphics 28*, 1 (2021), 151–161. 1

[IS18] Iglovikov V., Shvets A.: TernausNet: U-Net with VGG11 Encoder Pre-Trained on ImageNet for Image Segmentation. *ArXiv e-prints* (2018). arXiv:1801.05746. 2, 3

[IZ18] Islam J., Zhang Y.: Towards robust lung segmentation in chest radiographs with deep learning. *arXiv preprint arXiv:1811.12638* (2018). 2, 3

[LHL*20] Li Y., Horowitz M. A., Liu J., Lan H., Liu Q., Sha D., Yang C.: Individual-level fatality prediction of COVID-19 patients using AI methods. *Frontiers in Public Health 8* (2020), 566. 1, 4

[PMR*20] Ponti G., Maccaferri M., Ruini C., Tomasi A., Ozben T.: Biomarkers associated with COVID-19 disease progression. *Critical reviews in clinical laboratory sciences 57*, 6 (2020), 389–399. 2, 3

[RM05] Rokach L., Maimon O.: Clustering methods. In *Data mining and knowledge discovery handbook*. Springer, 2005, pp. 321–352. 3

[SSP*21] Saltz J., Saltz M., Prasanna P., Moffitt R., Hajagos J., Bremer E., Balsamo J., Kurc T.: Stony Brook University COVID-19 Positive Cases (COVID-19-NY-SBU). The Cancer Imaging Archive, 2021. URL: https://doi.org/10.7937/TCIA.BBAG-2923. 1, 2

[VB18] Van Buuren S.: *Flexible imputation of missing data*. CRC Press, 2018. 3

[VdMH08] Van der Maaten L., Hinton G.: Visualizing data using t-SNE. *Journal of machine learning research 9*, 11 (2008). 3

[VLB*17] Valindria V. V., Lavdas I., Bai W., Kamnitsas K., Aboagye E. O., Rockall A. G., Rueckert D., Glocker B.: Reverse classification accuracy: Predicting segmentation performance in the absence of ground truth. *IEEE Transactions on Medical Imaging 36*, 8 (2017), 1597–1606. 2, 3

[ZCH*20] Zhao Z., Chen A., Hou W., Graham J. M., Li H., Richman P. S., Thode H. C., Singer A. J., Duong T. Q.: Prediction model and risk scores of ICU admission and mortality in COVID-19. *PloS one 15*, 7 (2020), e0236618. 1, 2, 4

[ZR21] Zebin T., Rezvy S.: COVID-19 detection and disease progression visualization: Deep learning on chest X-rays for classification and coarse localization. *Applied Intelligence 51*, 2 (2021), 1010–1021. 2

[ZVA*20] Zwanenburg A., Vallières M., Abdalah M. A., Aerts H. J. W. L., et al.: The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology 295*, 2 (2020), 328–338. 2