Big Data or Good Data: Which One Is More Important for AI in Medical Imaging?

Ran Zhang[1], Xin Tie[1], Zhihua Qi[3], Nicholas Bevins[3], John W. Garrett[2] and Guang-Hong Chen[1,2]

[1] Department of Medical Physics, University of Wisconsin-Madison, Madison, WI, USA
[2] Department of Radiology, University of Wisconsin-Madison, Madison, WI, USA
[3] Department of Radiology, Henry Ford Health

Failures of AI in the COVID Pandemic

ARTIFICIAL INTELLIGENCE

Hundreds of AI tools have been built to catch covid. None of them helped.

Some have been used in hospitals, despite not being properly tested. But the pandemic could help make medical AI better.

By Will Douglas Heaven July 30, 2021

Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans

Michael Roberts, Derek Driggs, Matthew Thorpe, Julian Gilbey, Michael Yeung, Stephan Ursprung, Angelica I. Aviles-Rivero, Christian Etmann, Cathal McCague, Lucian Beer, Jonathan R. Weir-McCall, Zhongzhao Teng, Effrossyni Gkrania-Klotsas, AIX-COVNET, James H. F. Rudd, Evis Sala & Carola-Bibiane Schönlieb

Nature Machine Intelligence **3**, 199–217 (2021) | Cite this article

Why AI Failed to Live Up to Its Potential During the Pandemic

by Bhaskar Chakravorti

March 17, 2022

AI for radiographic COVID-19 detection selects shortcuts over signal

Alex J. DeGrave, Joseph D. Janizek & Su-In Lee

Nature Machine Intelligence **3**, 610–619 (2021) | Cite this article

The Fundamental Challenge: Generalizability

- Generalizability in the context of statistical learning
  - Consistent performance on the i.i.d test set
  - Pitfalls (bias, shortcuts)

- Generalizability in the context of medical AI
  - Consistent performance on prospective, external clinical cohorts (i.i.d assumption may not be valid)
  - Learn the desired solution that generalizes to the target cohort

- Ways to improve generalizability?
  - Data size
  - Data heterogeneity

Background: COVID Classification from CXR

- Can a model trained using a small, high-quality dataset from a single clinical site be generalizable?
- How do the model's performance and generalization depend on the data size?

Data curation with quality assurance

- All data are collected in the native DICOM format
- Metadata such as patient sex, patient age, viewpoint, modality, imaging system vendor, and model are collected to check for potential biases
- A short time window (-3 to 3 days) between the imaging study and RT-PCR test was used to ensure the accuracy of the diagnosis (label)
- Both COVID+/COVID- cohorts were collected from the same hospitals and within the same time range to mitigate shortcut learning

Datasets for model training and evaluation

Model development

	HF-train
Type	
Time	Feb-Sep, 2020
No. images (+/-)	6689/10848
No. patients (+/-)	3264/4802
Age (+/-)	63±17/69±15
Imaging system vendors	Carestream (53%), Konica Minolta (20%), GE (19%), Agfa (6%), others (2%)

https://bimcv.cipf.es/bimcv-projects/bimcv-covid19/

HENRY FORD HEALTH

BIMCV
Medical Imaging Databank of the Valencia Region

25,000 CXRs from 15,000 patients

UW Health

MIDRC
MEDICAL IMAGING AND DATA RESOURCE CENTER.

https://www.midrc.org/

Sampled training datasets with different sizes

- Data size from 100 patients to 6000 patients
  - 50/50 class ratio
  - 10 different random samples for each data size
  - For each data size, 10 different models are trained. The mean and the standard deviation of AUC are calculated
- Evaluation of generalizability
  - AUC gap between internal test and external tests

Model description

- Model architecture: DenseNet-121[1]
- Three-stage transfer learning
  - ImageNet dataset
  - NIH chest x-ray dataset[2]
  - COVID CXR dataset
- Model ensemble
  - Five models trained with different Train/Val splits

1. Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely connected convolution-al networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, July 21–26, 2017.
2. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2017). Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2097-2106).

AUC vs training data size

	Internal	BIMCV	UW Health	MIDRC
100	0.732 ±0.016	0.725 ±0.023	0.729 ±0.024	0.700 ±0.025
200	0.766 ±0.026	0.749 ±0.028	0.764 ±0.018	0.731 ±0.019
400	0.786 ±0.009	0.772 ±0.015	0.779 ±0.009	0.747 ±0.010
800	0.794 ±0.007	0.781 ±0.009	0.785 ±0.006	0.757 ±0.009
1200	0.799 ±0.008	0.787 ±0.007	0.791 ±0.005	0.764 ±0.009
1600	0.802 ±0.003	0.792 ±0.005	0.797 ±0.006	0.766 ±0.005
2000	0.808 ±0.004	0.796 ±0.005	0.800 ±0.005	0.771 ±0.006

$$\text{AUC} = aN^k + b$$

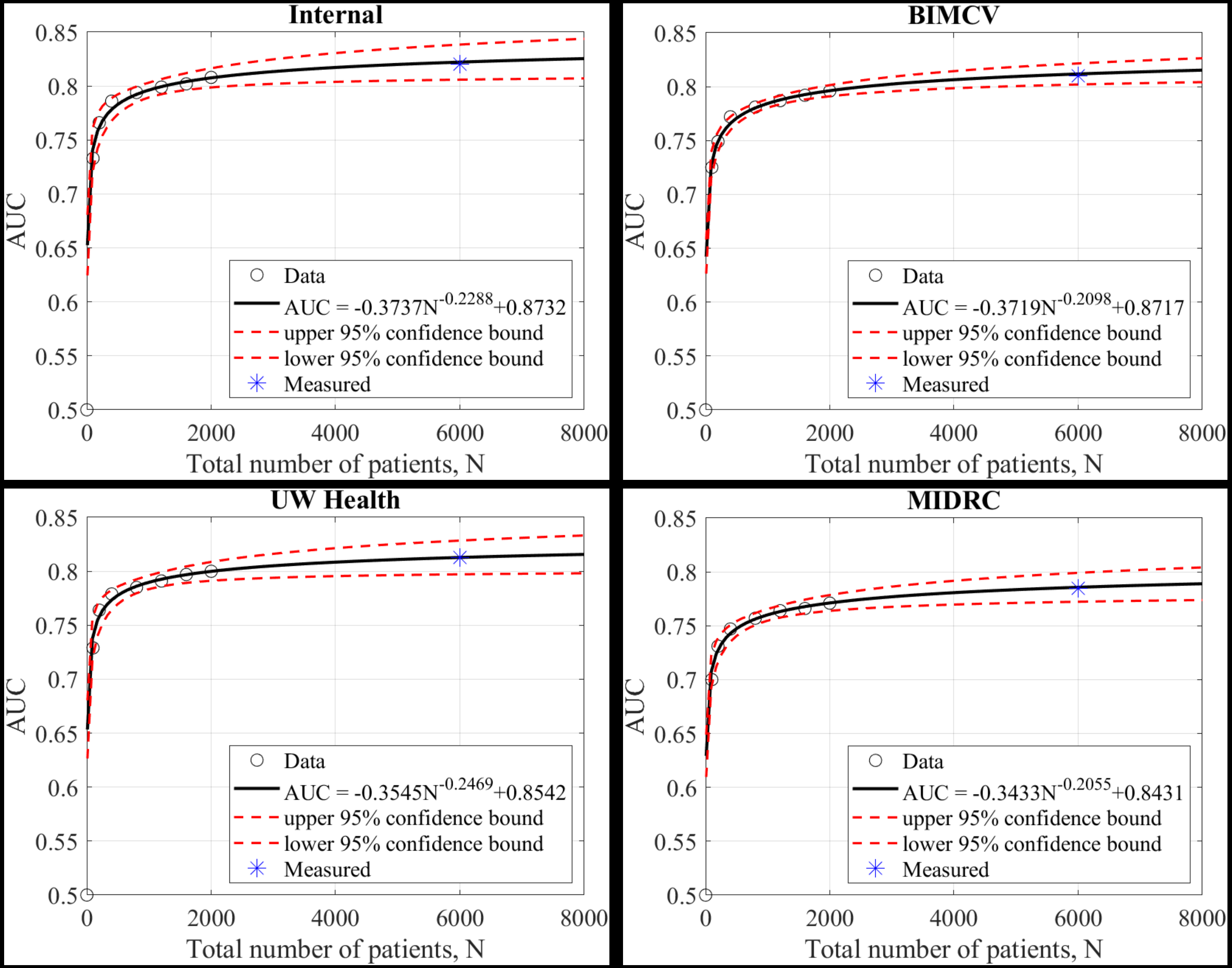
AUC vs training data size

	Internal	BIMCV	UW Health	MIDRC
100	0.732 ±0.016	0.725 ±0.023	0.729 ±0.024	0.700 ±0.025
200	0.766 ±0.026	0.749 ±0.028	0.764 ±0.018	0.731 ±0.019
400	0.786 ±0.009	0.772 ±0.015	0.779 ±0.009	0.747 ±0.010
800	0.794 ±0.007	0.781 ±0.009	0.785 ±0.006	0.757 ±0.009
1200	0.799 ±0.008	0.787 ±0.007	0.791 ±0.005	0.764 ±0.009
1600	0.802 ±0.003	0.792 ±0.005	0.797 ±0.006	0.766 ±0.005
2000	0.808 ±0.004	0.796 ±0.005	0.800 ±0.005	0.771 ±0.006
Learning curve	$-0.3755N^{-0.2191} + 0.8751$	$-0.3720N^{-0.2079} + 0.8718$	$-0.3529N^{-0.2521} + 0.8525$	$-0.3433N^{-0.2055} + 0.8431$

AUC vs training data size

	Internal	BIMCV	UW Health	MIDRC
100	0.732 ±0.016	0.725 ±0.023	0.729 ±0.024	0.700 ±0.025
200	0.766 ±0.026	0.749 ±0.028	0.764 ±0.018	0.731 ±0.019
400	0.786 ±0.009	0.772 ±0.015	0.779 ±0.009	0.747 ±0.010
800	0.794 ±0.007	0.781 ±0.009	0.785 ±0.006	0.757 ±0.009
1200	0.799 ±0.008	0.787 ±0.007	0.791 ±0.005	0.764 ±0.009
1600	0.802 ±0.003	0.792 ±0.005	0.797 ±0.006	0.766 ±0.005
2000	0.808 ±0.004	0.796 ±0.005	0.800 ±0.005	0.771 ±0.006
Learning curve	$-0.3755N^{-0.2191} + 0.8751$	$-0.3720N^{-0.2079} + 0.8718$	$-0.3529N^{-0.2521} + 0.8525$	$-0.3433N^{-0.2055} + 0.8431$
6000 (prediction)	0.819	0.811	0.813	0.786
6000 (measured)	0.818	0.809	0.813	0.786

AUC vs training data size



- To boost AUC from 0.81 to 0.83 on the UW dataset requires an increase of N from 6,000 to 100,000
- Small training datasets (~100 patients) can be used to develop a baseline model with good initial performance

Important lessons learned

- Data quality >> data size
- Model trained using well-curated data from a single clinical site can generalize to other sites
- A small, high-quality training dataset can provide a decent baseline

Thank You

University of Wisconsin-Madison