UNIVERSITY OF CAMBRIDGE

# Incorporation Bias in Medical Machine Learning Models

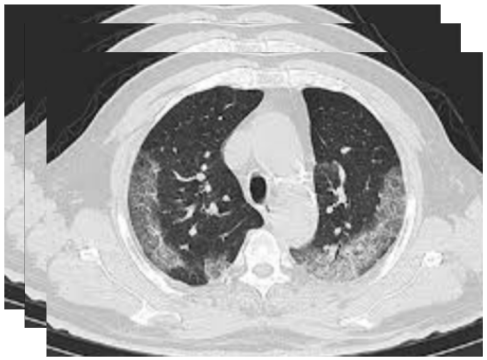
Derek Driggs, Presenting at the Isaac Newton Institute

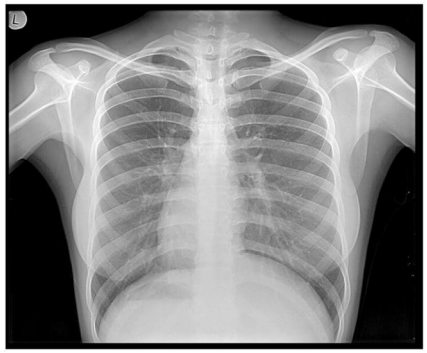
# The AIX-CovNet Collaboration

- **Investigators**: Carola-Bibiane Schönlieb (PI), Evis Sala (PI), Judith Babar, Anna Korhonen, Lorena Escudero, Mishal Patel, **Michael Roberts**, James Rudd, Alessandro Ruggiero, Zhongzhao Teng, Muhunthan Thillai.
- **AI and image analysis team**: Angelica Aviles-Rivero, ***Derek Driggs***, Christian Etmann, Julian Gilbey, Paula Martin Gonzales, Johannes Hofmanninger, Emily Jefferson, Georg Langs, Pietro Lio, Jan Stanczuk, Phil Teare, Matthew Thorpe, Jing Tang, Nicholas Walton, Guang Yang, Michael Yeung, Xiaoxiang Zhu.
- **Clinical team:** Emmanuel Ako, Lucian Beer, Effrossyni Gkrania-Klotsas, Cathal McCague, Jacobus Preller, Helmut Prosch, Ian Selby, Jonathan Weir-McCall, Kang Zhang.
- **Data contributors:** Addenbrooke's Hospital, Royal Papworth Hospital, Chelsea and Westminster Hospital, Hubei Renmin Hospital, Tongji Hospital, Beijing Youan Hospital, Luohe Centre Hospital, NHSX / British Society for Thoracic Imaging, Medical University Vienna.

# Our Ambitions

- Using quality-controlled longitudinal radiological imaging and clinical data to predict patient outcomes for **resource management**





**Clinical Measurements**

- Age
- Sex
- Temperature
- Time since symptoms start
- SpO2 / FiO2
- D-dimer
- Comorbidities
- CRP
- ….

Oxygen support

NEWS Score

Death

Respiratory Failure

# Advice from Prof. Gog in April

"Rather than adding noise, **amplify the signal**."

"Although your instinct may be to start your own models from scratch…."

## How you can help with COVID-19 modelling

*Julia R. Gog*

Many physicists want to use their mathematical modelling skills to study the COVID-19 pandemic. Julia Gog, a mathematical epidemiologist, explains some ways to contribute.

Credit: Marisa Crimlis-Brown

While the COVID-19 pandemic continues its global devastation, the instinctive reaction from scientists is "how can we help?" I will try to answer this in general terms for colleagues with expertise in mathematics and modelling, but who may have little or no prior experience with infectious disease modelling.

Clearly, the set of things that would not help includes rediscovering results that disease modellers have known for decades — or for more than a century in the case of

**Communicating to the public**

The world wants to know what the science is behind the decisions, but there is great danger of misinformation when media interest is amplifying the voices of scientists, but not necessarily those most qualified to comment. You can learn the mathematical and scientific ideas from the broader literature, including some great textbooks. (Real-time papers are aimed at colleagues who know the literature already; reading only these will not be enough

# One Year Later, Have We Been Effective?

Thousands of papers develop a machine learning model for COVID-19 diagnosis or prognostication. **Are they trustworthy?**

At least two models have been deployed in China or in Europe [1]. **What qualities of these studies make the models usable?**

How should ML researchers **continue to contribute, now and in future global crises?**

[1] Wang, Minghuan, et al., “Deep learning-based triage and analysis of lesion burden for COVID-19: a retrospective study with external validation”. *The Lancet* (2020).

# Our Systematic Review Investigates This Question [2]

29 papers had sufficient documentation to be **reproducible** (CLAIM/RQS cutoff)

8/29 use **severely biased datasets** (more on this later)

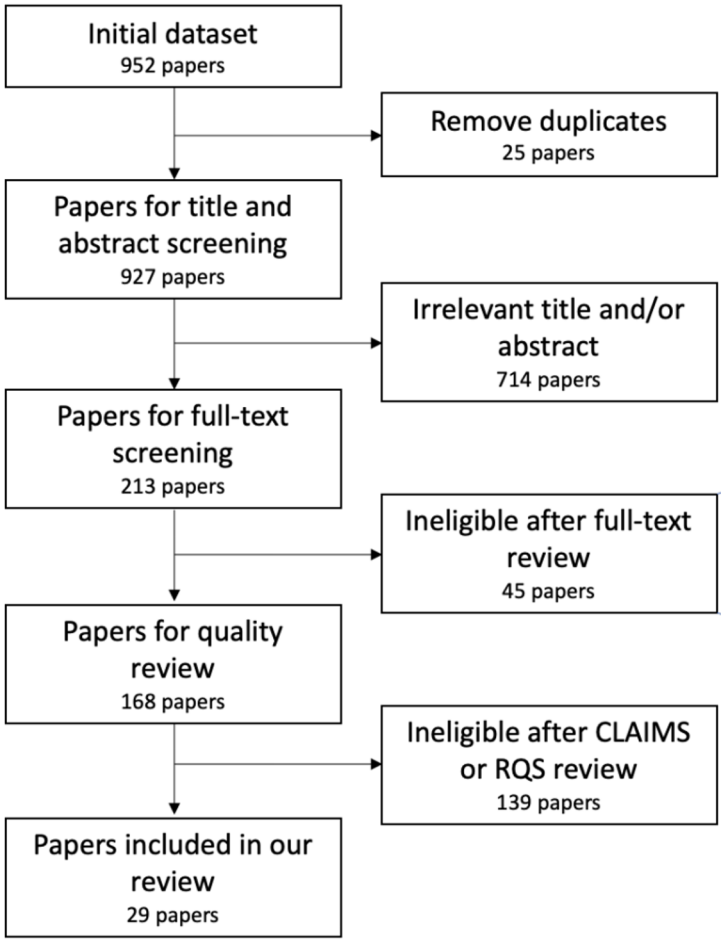
Reliance on “Frankenstein” public datasets

6/29 share their code and model

23/29 have **no external validation**

16/29 have **no sensitivity or robustness analysis**

**None of these models are clinically useful**




[2] Roberts, M., Driggs, D., Thorpe, M., *et al.* "Machine learning for COVID-19 detection and prognostication using chest radiographs and CT scans: a systematic methodological review." *Nature Machine Intelligence* (2021).

# We Review Papers Against State-of-the-Art Medical Standards: CLAIM, RQS, PROBAST

*Table 2.* PROBAST: Summary of Step 3—Assessment of Risk of Bias and Concerns Regarding Applicability*

| 1. Participants | 2. Predictors | 3. Outcome | 4. Analysis |
|---|---|---|---|
| **Signaling questions** | | | |
| 1.1. Were appropriate data sources used, e.g., cohort, RCT, or nested case–control study data? | 2.1. Were predictors defined and assessed in a similar way for all participants? | 3.1. Was the outcome determined appropriately? | 4.1. Were there a reasonable number of participants with the outcome? |
| 1.2. Were all inclusions and exclusions of participants appropriate? | 2.2. Were predictor assessments made without knowledge of outcome data? | 3.2. Was a prespecified or standard outcome definition used? | 4.2. Were continuous and categorical predictors handled appropriately? |
| – | 2.3. Are all predictors available at the time the model is intended to be used? | 3.3. Were predictors excluded from the outcome definition? | 4.3. Were all enrolled participants included in the analysis? |
| – | – | 3.4. Was the outcome defined and determined in a similar way for all participants? | 4.4. Were participants with missing data handled appropriately? |
| – | – | 3.5. Was the outcome determined without knowledge of predictor information? | 4.5. Was selection of predictors based on univariable analysis avoided?† |
| – | – | 3.6. Was the time interval between predictor assessment and outcome determination appropriate? | 4.6. Were complexities in the data (e.g., censoring, competing risks, sampling of control participants) accounted for appropriately? |
| – | – | – | 4.7. Were relevant model performance measures evaluated appropriately? |
| – | – | – | 4.8. Were model overfitting, underfitting, and optimism in model performance accounted for?† |
| – | – | – | 4.9. Do predictors and their assigned weights in the final model correspond to the results from the reported multivariable analysis?† |

# The State of the Art at the Pandemic's Beginning

CheXNet achieves ***superhuman performance*** on pneumonia detection task

**Model**: 121-layer DenseNet

**Data**: 250,000 chest radiographs with many pathologies

**Training Labels**: Extracted from radiology reports

**Testing Labels**: Consensus of four radiologists

"**Human-Machine** agreement is higher than average **Human-Human** agreement."

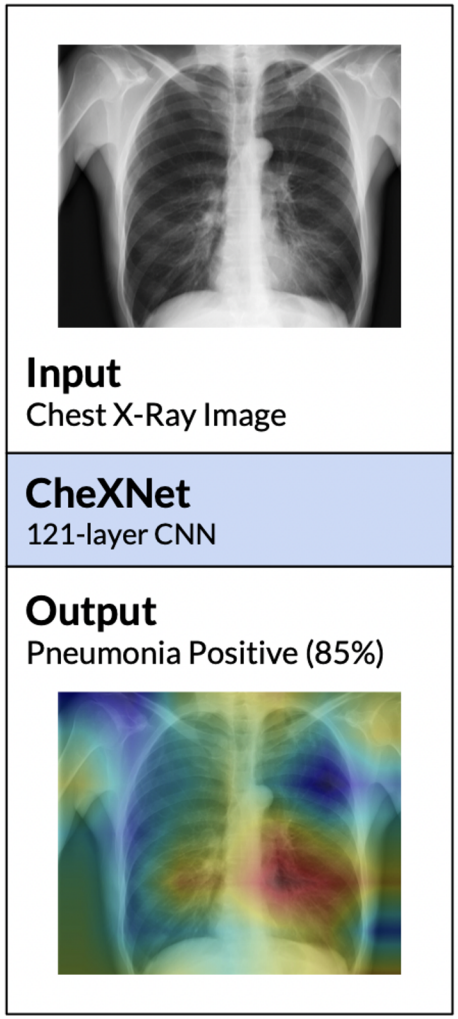



Figure from Rajpurkar, P., *et al.*, "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning", *arXiv:1711.05225*. (2017).

# Interpreting the Strong Results from CheXNet

"**Human-Machine** agreement is higher than average **Human-Human** agreement."

Paradox: **higher-variance labels** likely lead to **better performance**

The problem is bigger than "noisy labels"

If committee **perfectly agreed**:

- Superhuman performance would be **impossible**
- Performance **would not generalise**

# Incorporation Bias

If **labels are not independent of the predictors**, the model can suffer from **incorporation bias**

Incorporation bias almost always yields optimistic performance metrics

For example: COVID-19 labels determined from imaging features

Models depending on **imaging data** or **subjective features** are particularly vulnerable [3]

This bias is a **spectrum**:

Example: Patient outcome > Patient treatment > Radiologist report

[3] Moons, K. et al., “PROBAST: A tool to assess risk of bias and applicability of prediction model studies”. Ann. Intern. Med. (2019).

# Examples of Incorporation Bias

COVID-19 labels determined from imaging features [4]

**Human-in-the-loop** segmentation strategies

Using the full data set to train a **GAN for data augmentation**

It is potentially inappropriate to train to **replicate radiologists**

[4] Wang, G. *et al.*, “A deep-learning pipeline for the diagnosis and discrimination of viral, non-viral and COVID-19 pneumonia from chest X-ray images”. *Nat. Bio. Eng.* (2021).

# Accounting for Incorporation Bias

If no **gold standard** exists, incorporation bias is almost unavoidable

Using **more predictors** and **more tests** can limit the effects

**Open Problem: quantify** model performance given labeller differences

# Conclusion

**Incorporation bias** is prevalent in medical machine learning models

Many **recognised generalisation issues stem from incorporation bias**

Studies exhibiting incorporation bias can still be **high-quality** and **useful**

**More research is necessary** to understand the effects of incorporation bias

# Questions?

We are happy to collaborate with:

- **clinicians**, **data contributors**, groups that can augment our algorithmic approaches.

Website: https://covid19ai.maths.cam.ac.uk/

**Review:** “Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans”. *Nat. Mach. Intel.*

**Invited Editorial:** “Machine Learning for COVID-19 Diagnosis and Prognostication: Lessons for Amplifying the Signal While Reducing the Noise”. *Rad. A.I.*


email: dtd24@cam.ac.uk