



PHILIPS

# Challenges of using medical datasets

Dennis Moolenaar, Diana Cojocaru

\_Business

July 2nd 2021

innovation  you

# AGENDA

- Introduction
- What kind of medical datasets are available?
- Challenges
- Data Quality
- Conclusions

It is our purpose to improve people's health and well-being through meaningful innovation.

We aim to improve the lives of 2.5 billion people per year by 2030\*

\* 2 billion by 2025



# How do we improve lives ?

- Create meaningful innovation
- These innovations can require Artificial Intelligence
- To create AI we need access to datasets
- To access data we need
  - Check License
  - Follow Privacy Process

# What kind of medical datasets are available?

- Primary data collection
  - information obtained directly from the source
- Retrospective data usage
  - existing data that was previous recorded
- Synthetic data generation
  - generated to represent the authentic data
  - data is artificially created

# License



- Check license type of the website where the data is located
- Commercial companies cannot use data that's under CC non-commercial license
- Non-commercial companies have more flexibility and can easily access public data
- Check with a legal expert



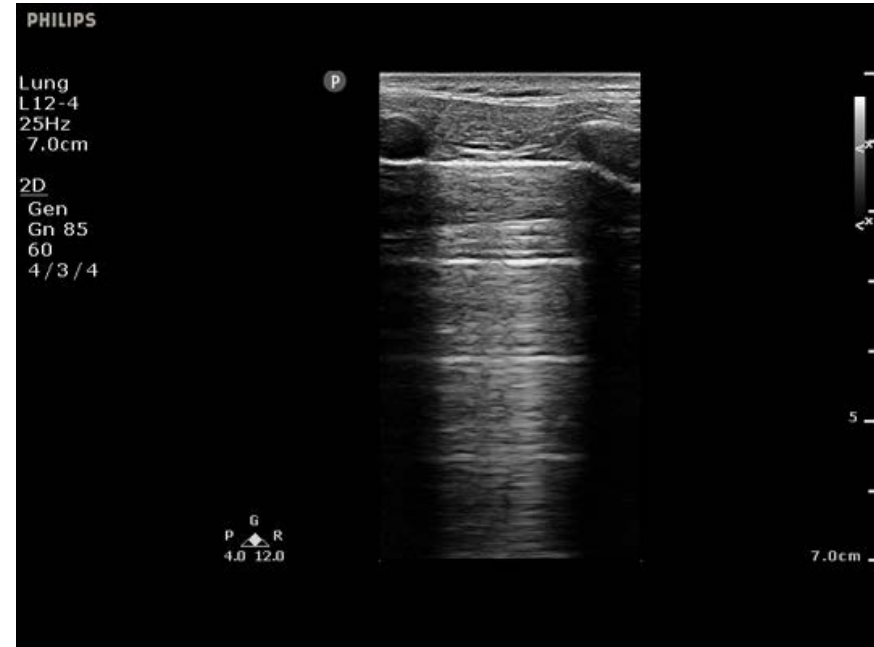


# Philips Privacy Process

- Philips is compliant with privacy laws
- Privacy laws are applicable if data can be linked to a person
- Privacy laws are very complex
- Always involve privacy experts early ( and throughout )
- Philips implements robust procedures
  - To carry out data protection impact assessments
  - Implements Philips privacy rules
  - Implements Philips binding corporate rules

# Our Use Case: Detection of Healthy/Unhealthy lungs

- Detect Healthy/Unhealthy lungs from Lung Ultrasounds images
  - Unhealthy is Pneumonia/Covid-19
- Synthetic Data Generation is not an option
- Images are anonymized => No problem ?





# Privacy Pitfalls

- Anonymized data is not always anonymized
  - Combination of datasets can link data to person
  - Rare medical conditions can reveal identity
- Synthetic data can be personal data
  - If synthetic data is based on real data and can be linked

# Data quality

- One of the reasons AI projects thrive or fail
- Completeness, correctness, and appropriateness of the data (video quality)
- Data annotation
  - Indispensable stage in developing neural network algorithms
  - Data must be properly categorized and labeled by experts in the field
- Model performance can degrade if quality real data is different from training data

# CONCLUSIONS

- To improve lives, we need quality (medical) data
  - Quality Medical data is vital for the development of accurate neural networks
- But license and privacy laws need to be taken into account
- AI poses challenges for privacy



Thank you

