Appling AI To 2D Medical Imaging

Intuition About 2D Medical Images

Different Clinical Imaging Tools

X-Rays	Computed Tomography	Magnetic Resonance Imaging
$	$$	$$$
Radiation	X-Rays with computer technology	Strong magnetic fields
Not much detail	More details	More details
Bone, lungs, heart, imaging patients with metal	Organs, soft tissue, muscles, lung and chest, imaging patients with metal	Joints, brain, organs, soft tissue an internal structures

x-ray

This type of imaging projects a type of radiation called x-rays down at the body from a single direction to capture a single image. These are relatively

cheap to acquire by imaging standards, but the downside is that they emit radiation and don’t offer too much detail at the organ-level. They allow us to see major structures like bone, lungs, and heart, and they’re safe to use for people who have metal in their bodies.

Computed Tomography (CT)

CT uses x-ray, but they emit x-rays from many different angles around the human body to capture more detail from more different angles. They are more expensive, but they allow us to see more details about organs and soft tissues in the body. Most hospitals in the US have a CT scanner in them.

Magnetic Resonance Imaging (MRI)

MRI uses strong magnetic fields and radio waves to create images of areas of the body from all different angles. It allows us to assess even more details about the human body. This type of imaging is particularly useful for studying the human brain. Although it is safer in that it does not emit radiation, MRI is not safe for people with metal in their bodies. Not all hospitals in the US have MRI scanners and it is a very expensive imaging tool.

Ultrasound

Ultrasound isn't covered in the video. It utilizes high-frequency sound waves beyond the audible limit of human hearing to generate images. Ultrasound waves travel through soft tissues or fluids and bounce back when it hits dense tissues. More waves bounce back if the tissue is denser. The waves that bounce back are captured to generate images. Ultrasound is very safe and commonly used during pregnancy.

2D vs. 3D

Out of these different imaging tools, x-ray is the **only** 2D imaging tool.

2D imaging takes the picture from a single angle, and everything that is visible to the device at that angle will appear in the picture. You do not see overlapping structures in the 2D image.

3D imaging takes lots of pictures from lots of angles to create a **volume** of images. You can see structures that are behind one another. The final 3D image is actually a set of 2D images that represent different slices through the body. So, in any single slice or 2D image, you can't see the whole part but if you scroll through the slices, you will view the **volume** of that body part.

New terms

- **X-ray**: a 2D imaging technique that projects a type of radiation called x-rays down at the body from a single direction to capture a single image.
- **Ultrasound**: a 2D imaging technique that uses high-frequency sound waves to generate images.
- **Computed Tomography (CT)**: a 3D imaging technique that emits x-rays from many different angles around the human body to capture more detail from more different angles.
- **Magnetic Resonance Imaging (MRI)**: a 3D imaging technique that uses strong magnetic fields and radio waves to create images of areas of the body from all different angles.
- **2D imaging**: an imaging technique that pictures are taken from a single angle.
- **3D imaging**: an imaging technique that pictures are taken from different angles to create a volume of images.

Clinical Applications

Clinical Applications

Radiologist

The primary reader of medical imaging data is a type of clinician called a radiologist. These clinicians read all types of 2D and 3D images from all areas of the body. Their role in the clinical workflow is to *read imaging studies* and write *interpretations* of the images that can then be understood by other clinicians who are not experts in imaging.

Diagnosing Clinician

After the radiologist reads an imaging study, their radiology report is sent to the patient's *diagnosing clinician*. This clinician could be an emergency room doctor, primary care physician (PCP), or any other type of specialist. While the radiologist's report may have diagnostic information in it, the *final diagnosis* always comes from the diagnosing clinician who takes the radiologist's report into account alongside other information: the patient's medical history, lab results, and current symptoms. So, medical imaging plays a critical, but only a partial role in the diagnostic process.

Pathologists

Pathologists are a type of clinician who work primarily in *laboratories*. While radiologists are the primary readers of x-rays, CT, and MRI studies, pathologists are the primary readers of microscopy studies. It is their job to interpret findings from all different types of *cell-level* samples taken from patients such as tumor biopsies and blood smears.

Types of 2D Imaging

X-ray

The most common type of 2D imaging is x-ray. This technique uses a machine to emit x-rays, which are absorbed differently by different tissues in the body. Bone has *high absorption* and therefore appears *bright white*. Soft tissues like the heart and diaphragm absorb a *medium amount* and appear *gray*. Air *does not absorb* any x-rays and thus appears *black*.

We usually think of x-rays to look for fractures/broken bones, but two of their other most common use cases are for assessing abnormalities in the lungs, and for assessing breast tissue (mammograms).

Ultrasound

Ultrasound is a type of 2D imaging technique that isn't covered in the video. It utilizes high-frequency sound waves beyond the audible limit of human hearing to generate images. Ultrasound waves travel through soft tissues or fluids and bounce back when it hits dense tissues. More waves bounce back if the tissue is denser. The waves that bounce back are captured to generate images. Ultrasound is very safe and commonly used during pregnancy.

Microscopy

Microscopy refers to *physical slides* of biological material taken from a patient that can be viewed at the *cell-level* through a microscope. These slides often have a stain applied to them that causes different cell structures to appear in different colors. These stains help pathologists tell the difference between cell structures.

Fundal Imaging

The fundus of the eye is the interior surface of the eye, and images can be taken of it to diagnose diabetic retinopathy. In this condition, blood vessels at the back of the eye become damaged, so fundal imaging particularly looks at the integrity of the tiny vessels in the eye.

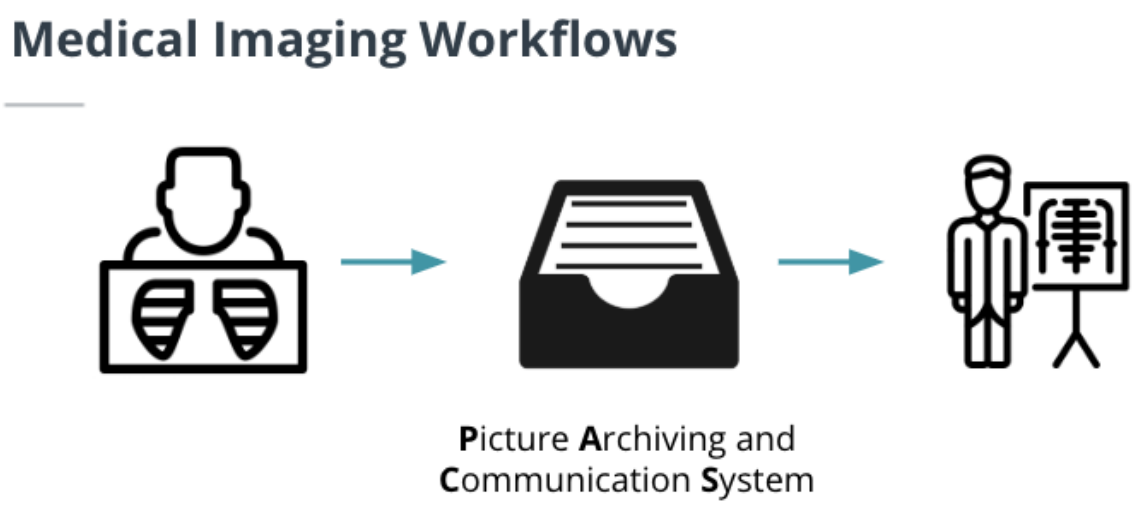
Differences in the imaging techniques

Since fundal images and microscopy images are not acquired with a digital machine, they are not inherently digital like x-rays are. As a result, an additional step of digitizing these images is required before applying AI. Once microscopy and fundal images are digitized, much of the AI principals can be applied to them the same way that they can be applied to x-ray images.

The second difference is that X-ray images are stored as single-channel grayscale images, while microscopy and fundal images are stored as red-green-blue (RGB) three-channel images.

Another major difference is that x-rays are stored in the DICOM format, which is the standard file format for medical imaging data, while this does not apply to microscopy and fundal images. We will cover more on this later.

Medical Imaging Workflows



Picture Archiving and Communication System (PACS)

Every imaging center and hospital have a PACS. These systems allow for all medical imaging to be stored in the hospital's servers and transferred to different departments throughout the hospital.

Diagnostic Imaging

In diagnostic situations, a clinician orders an imaging study because they believe that a disease *may be present* based on the patient's symptoms. Diagnostic imaging can be performed in *emergency* settings as well as *non-emergency* settings.

Screening Imaging

Screening studies are performed on populations of individuals who *fall into risk groups* for certain diseases. These tend to be diseases that are relatively common, have serious consequences, but also have the potential of being reversed if detected and treated early. For example, individuals who are above a certain age with a long smoking history are candidates for lung cancer screening which is performed using x-rays on an annual basis.

Types of 2D Imaging Algorithms

Classification

The classification algorithm assesses a whole image and returns an output stating *whether or not* a disease or abnormality is present in an image. These types of algorithms can be used for binary or multi-class classification, where

a single algorithm can classify for the presence or absence of multiple types of findings or diseases.

Localization

Localization algorithms are intended to aid radiologists in determining *where* in an image a particular finding is. These types of algorithms output a set of coordinates that create a *bounding box* around a section of the image where a particular type of finding is. These types of algorithms can be very useful for drawing radiologists' *attention* to certain types of findings that are difficult to see on imaging.

Segmentation

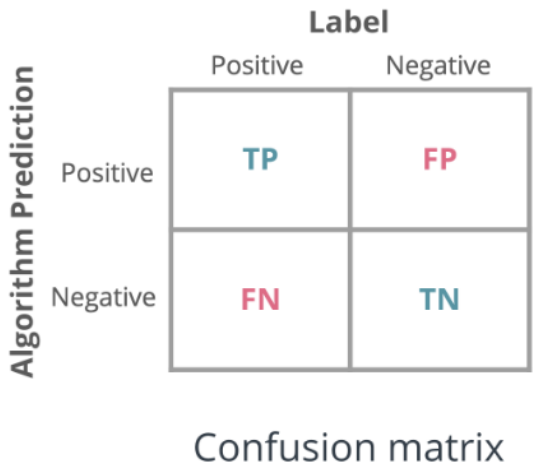
Segmentation algorithms return *a set of pixels* that contain the presence of a particular finding in an image, creating a *border* around a particular finding that allows for the calculation of its exact area. Segmentation algorithms are typically used to *measures the size* of particular findings or *count the number* of findings in an image. They are often used to count cells in microscopy data as well, where each cell in an image is segmented individually.

Clinical Impact of ML for 2D Imaging

You should be aware of the effect on *clinician workflows* when you are designing an algorithm that may be inserted into them.

Performance of ML

Label	Algorithm Prediction	
cancer (positive)	cancer (positive)	**True Positive**
cancer (positive)	no cancer (negative)	**False Negative**
no cancer (negative)	no cancer (negative)	**True Negative**
no cancer (negative)	cancer (positive)	**False Positive**



Confusion matrix

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

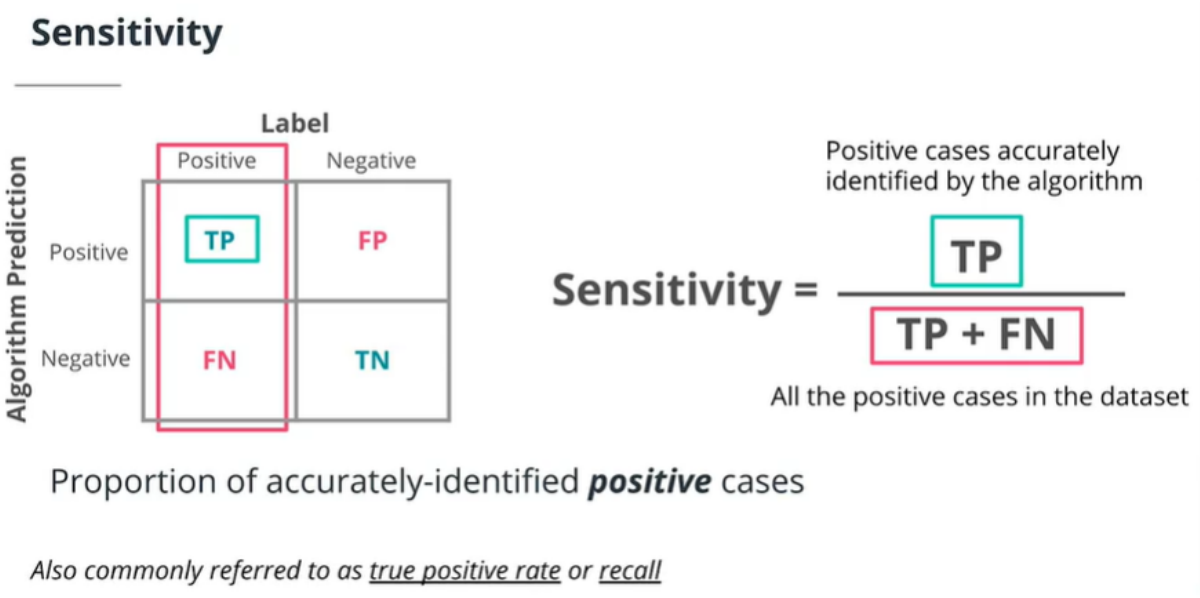
$$Dice(X, Y) = \frac{2*|X \bigcap Y|}{|X|+|Y|}$$

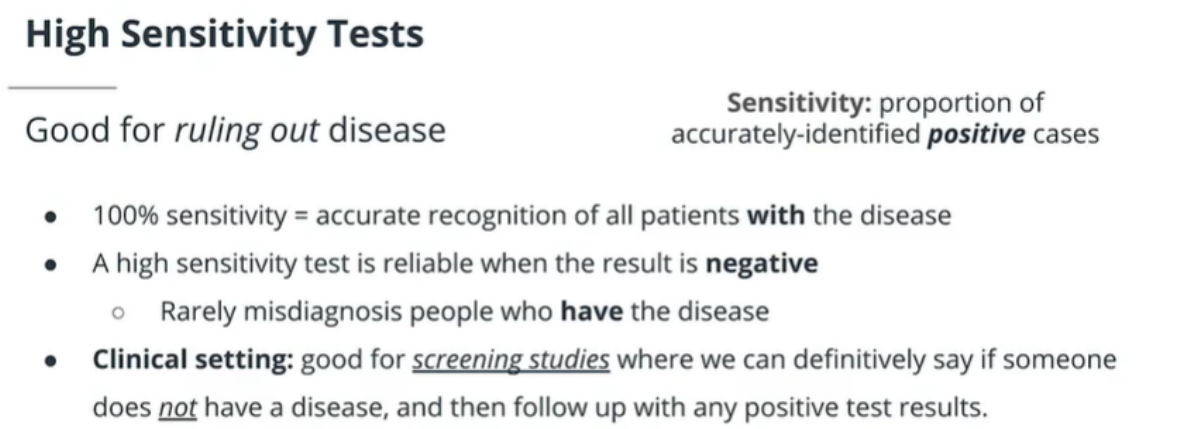
Performance Metrics

Sensitivity

Sensitivity is a metric that tells us among ALL the *positive* cases in the dataset, how many of them are successfully identified by the algorithm, i.e. the true positive. In other words, it measures the proportion of accurately-identified *positive cases*.

You can think of highly sensitive tests as being good for *ruling out* disease. If someone has a negative result on a highly sensitive algorithm, it is extremely likely that they don't have the disease since a high sensitive algorithm has low *false negative*.

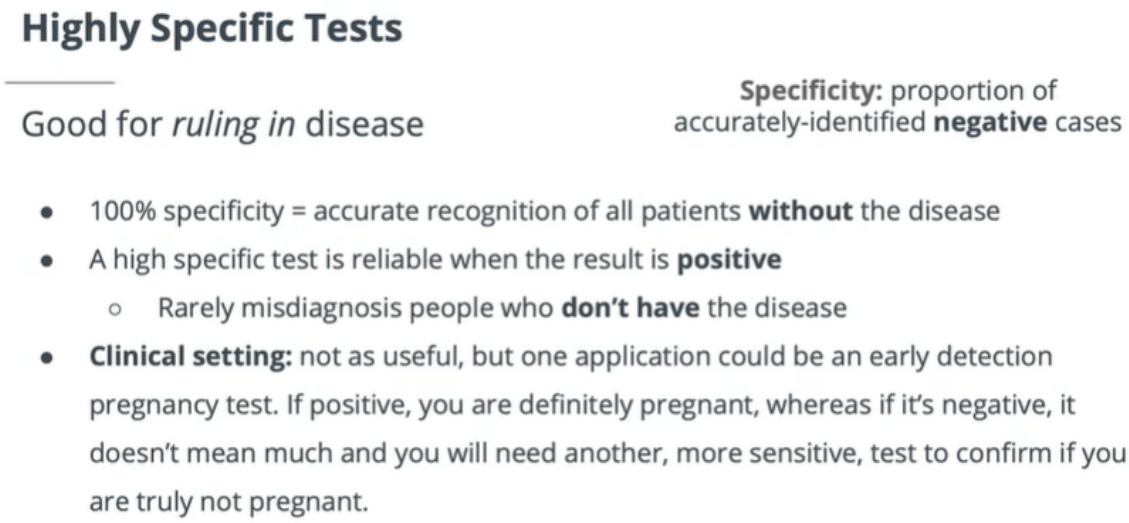




Specificity

Specificity measures ALL the *negative* cases in the dataset, how many of them are successfully identified by the algorithm, i.e. the true negatives. In other words, it measures the proportion of accurately-identified *negative* cases.

You can think of highly specific tests as being good for *ruling in* disease. If someone has a positive result on a highly specific test, it is extremely likely that they have the disease since a high specific algorithm has low *false positive*.



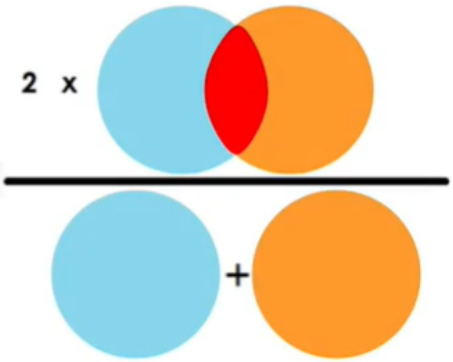
Dice coefficient

The dice coefficient measures the *overlap* of algorithm output and true labels. It is used to assess the performance of segmentation and localization.

Performance of Segmentation & Localization Algorithms

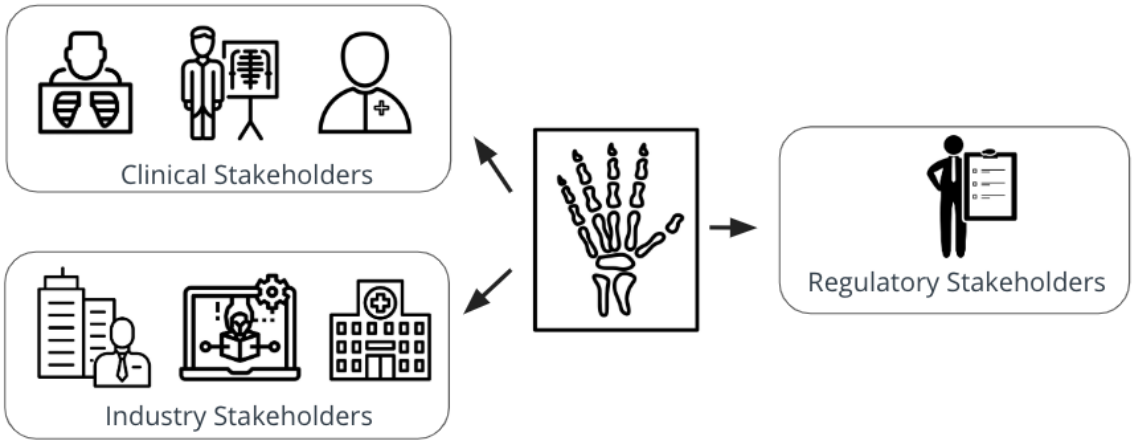
Performance is assessed by looking at overlap: **Dice Coefficient**

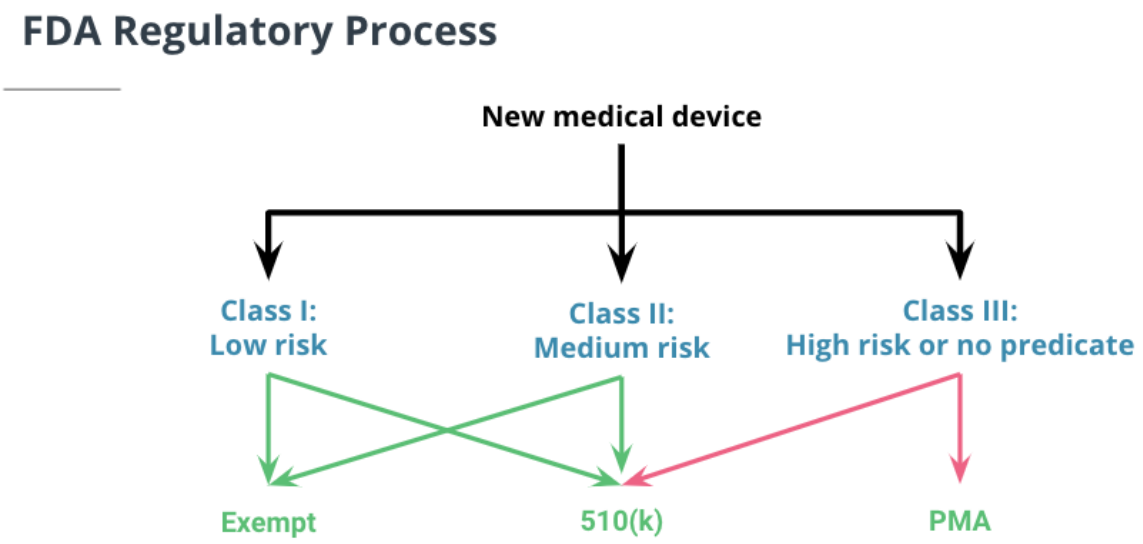
$$Dice(X,Y) = \frac{2*|X \bigcap Y|}{|X|+|Y|} =$$



Regulatory Landscape

Key Stakeholders





Key Stakeholders

Clinical Stakeholders

Clinical stakeholders are radiologists, diagnosing clinicians and patients. Radiologists are likely the end-users of an AI application for 2D imaging. They care about low disruption to workflow and they play an important advisory role in the algorithm development process. Clinicians have less visibility into the inner workings of an algorithm. They also care about low disruption to workflows and they care about the interpretability of algorithm output. Patients may be the most important stakeholder, and the FDA looks at your algorithm through the lens of protecting the patient from all unnecessary risks. Patients may never know that AI is involved and they care about the timeliness of receiving accurate test results.

Industry stakeholders

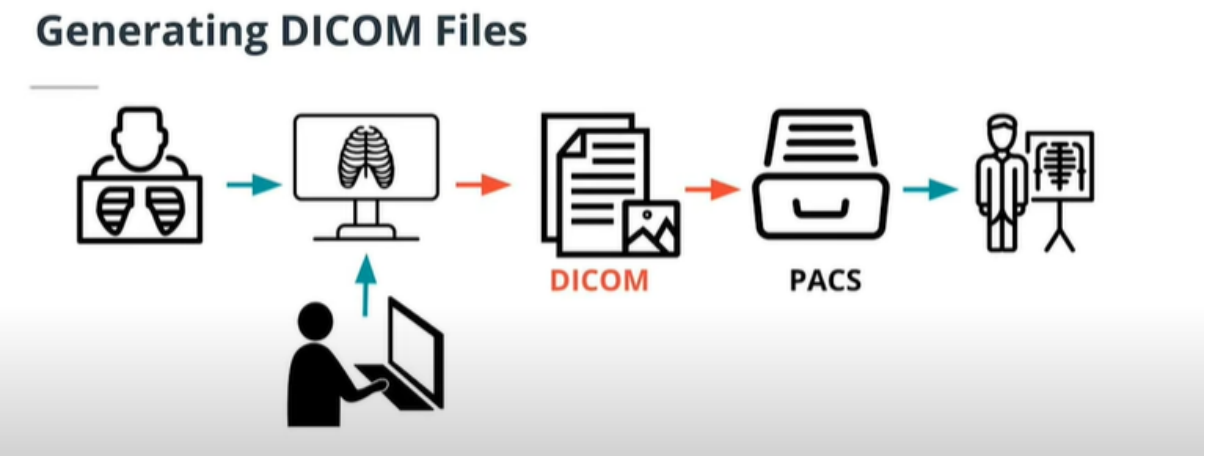
Industry stakeholders include medical device companies, software companies, and hospitals. Many medical device companies typically have accompanying imaging software. They also build their own AI algorithms to run on their hardware. Software companies can act more dynamically because they are not tied to a specific hardware system, but this also poses a regulatory challenge as the FDA wants to know if an algorithm performs the same across all hardware systems, and if not, which ones it is not appropriate for. Hospitals must be sure that they have the adequate infrastructure needed for algorithm deployment. In order to purchase an algorithm, a hospital must be convinced that it will save them money in the long run.

Regulatory stakeholder

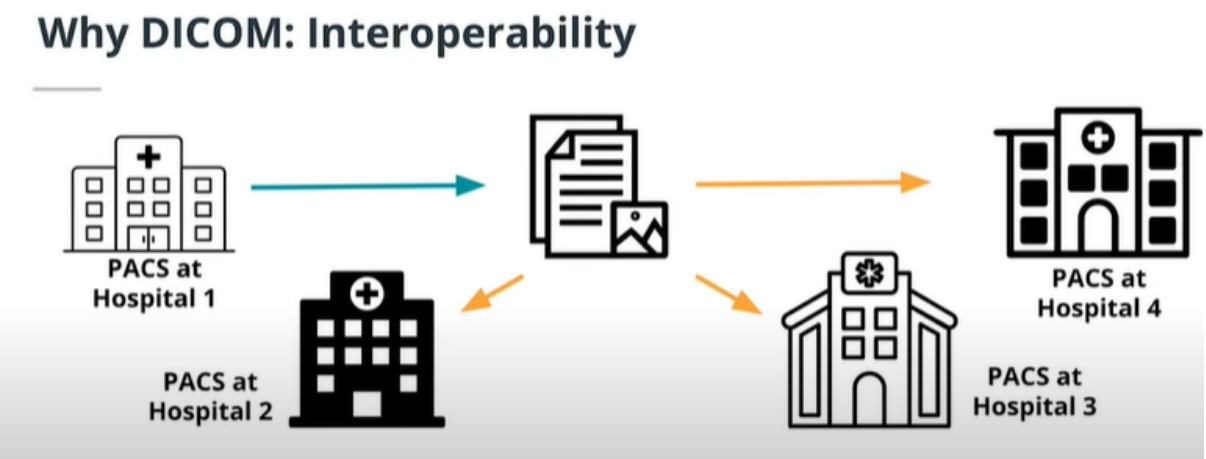
The main regulatory stakeholder in the medical imaging world is the Food and Drug Administration (FDA). The FDA treats AI algorithms as medical devices. Medical devices are broken down into three classes by the FDA, Class I, Class II, and Class III, based on their potential risks present to the patient. A device's class dictates the safety controls, which in turn dictates which regulatory pathway they must go down. The two main regulatory pathways for medical devices are **510(k)** and **Pre-market Approval (PMA).** Lower risk devices (Classes I & II) usually take a 501(k) submission pathway. Higher risk devices and algorithms (Class III) must go through PMA.

DICOM

DICOM is short for “Digital Imaging and Communications in Medicine”, which is the standard for the communication and management of medical imaging information and related data. DICOM files are a medical imaging file that is in the format that conforms to the DICOM standard.

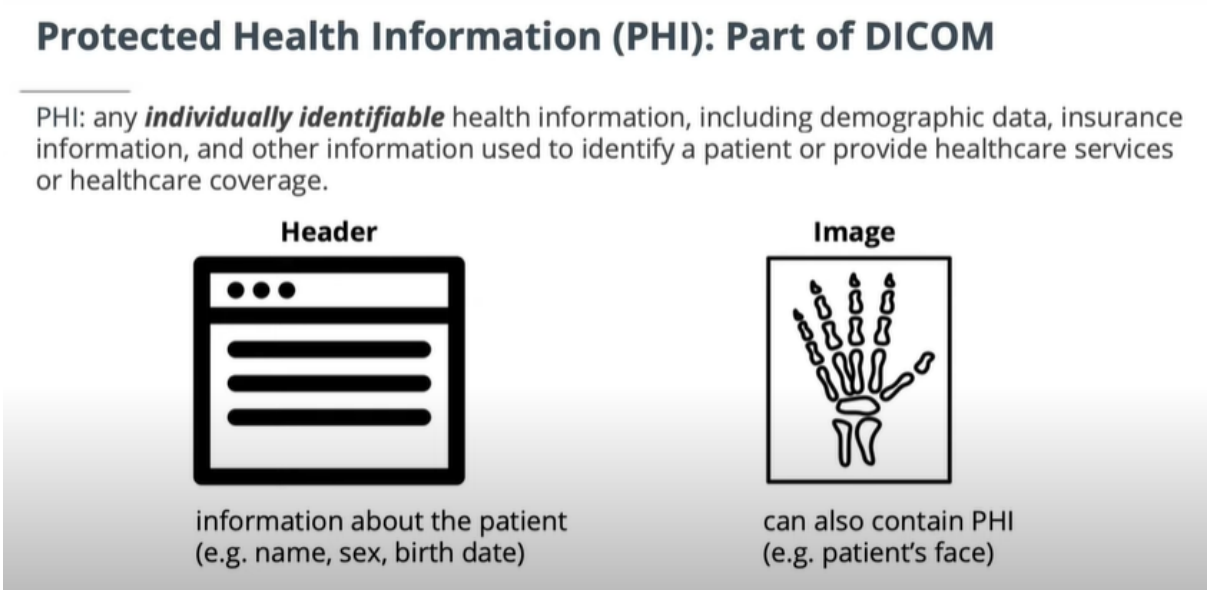


It was developed by the American College of Radiologists in 1993 to allow for interoperability.



A DICOM file contains information about the imaging acquisition method, the actual medical images, and patient information. It has a header component that contains information about the acquired image and an image component that is a set of pixel data representing the actual images

Protected Health Information (PHI) is part of DICOM and clinical data and radiologist report are *not* part of DICOM

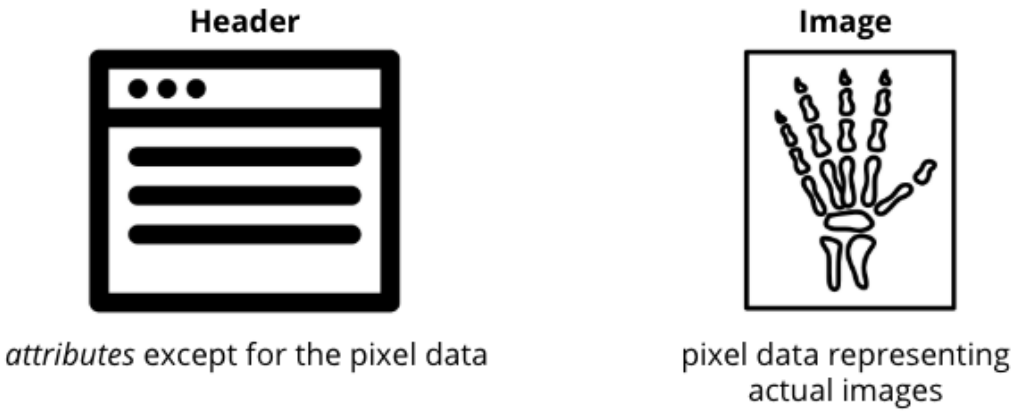


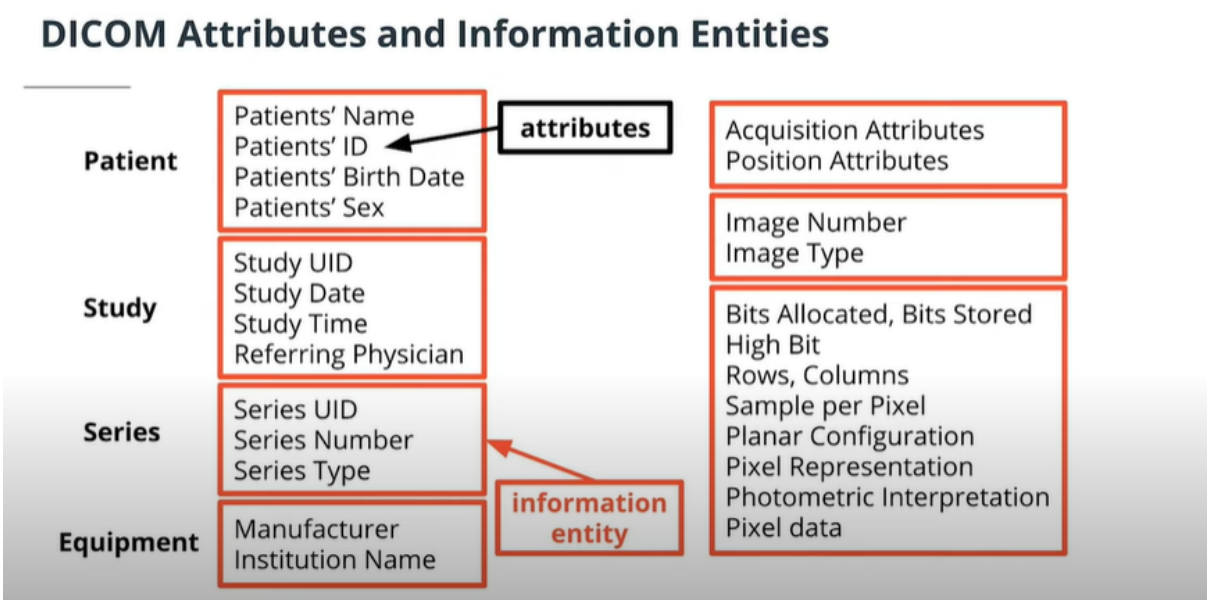
DICOM studies and series

DICOM Studies & Series

With 2D imaging, a single 2D image is known as a single DICOM series. All image series combined comprise a study of the patient, known as a DICOM study.

Components of a DICOM File





Read DICOM files

The pydicom package allows us to extract all of the metadata from a set of DICOM files and store them in a way that is easy to access.

```
my_dicom = pydicom.dcmread("MY_DICOM.dcm")
```

is used to read a DICOM file.

```
my_dicom_image = my_dicom.pixel_array
```

is used to access the actual pixel data of the image.

```
matplotlib.pyplot.imshow(my_dicom_image, cmap = 'gray')
```

is used to show the image.

It is worth noting that the data returned in `my_dicom.pixel_array` will be in the coordinate format `[y,x].`

Explore image data

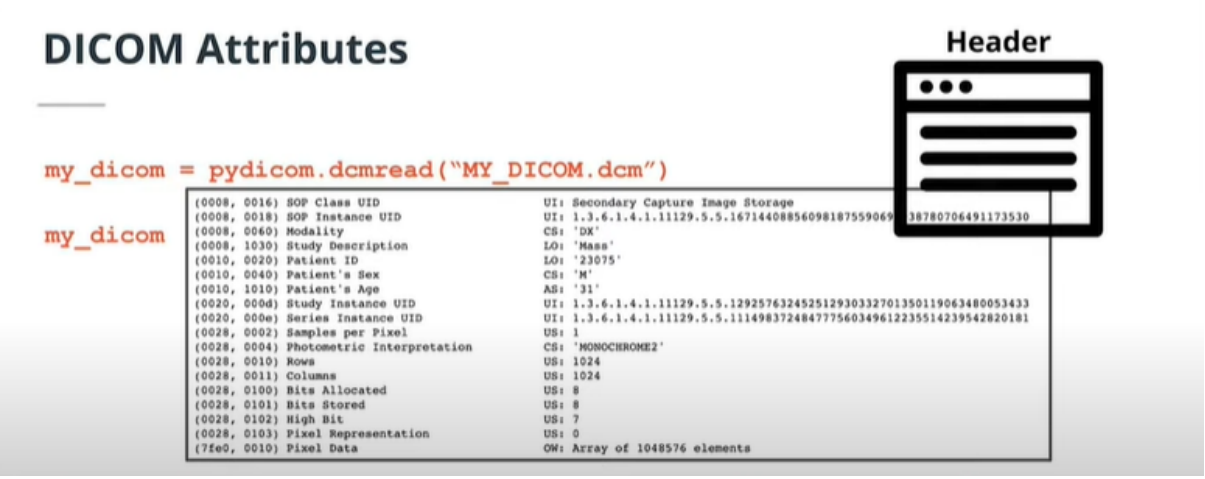
A good practice is to perform *random* spot checks of your data by choosing several *random* images and visualizing them. Then you can explore images in a

pixel level by looking at intensity profiles of individual images.

Prepare DICOM data

For efficient algorithm training, the best practice is to pre-extract all data from DICOM headers into a dataframe.

DICOM header has some other applications besides training models. It can be used to mitigate the risks of the algorithm. It can also be used to optimize image processing workflow.



Edge Case

Microscope to Digital pathalogy

Not all 2D medical images are stored as a DICOM. Microscopy images are not stored in DICOM since they do not come from a *digital* machine. Instead, they are biological data and come from smeared physical cells from patients.

The first step of transforming microscopy into a digital image is to get the cell sample from a patient. Then cells are dyed into different colors based on their structure and viewed by a microscope. The microscopy data is then captured by a camera to form a digital image. This transformation technique is called *digital pathology*.

Once images are digitized, they can be processed with ML in the same way as you would with the pixel data extracted from DICOM.

Classification Models of 2D Medical Images

ML vs. DL

The biggest difference between ML and DL is the concept of **feature selection.** Classical machine learning algorithms require predefined features in images. And, it takes up a lot of time and effort to design features. When deep learning came along, it was so groundbreaking because it worked to discover important features, taking this burden off of the algorithm researchers.

Ostu's method

It’s often used for background extraction and classification. It takes the intensity distribution of an image and searches it to find the intensity threshold that minimizes the variance in each of the two classes. Once it discovers that threshold, it considers every pixel on one side of that image to be one class and on the other side to be another class.

Convolutional neural network (CNN)

There are several sets of *convolutional layers* in a CNN model. Each layer is made up of a set of *filters* that are looking for features. Layers that come early in a CNN model look for very simple features such as directional lines and layers that come later look for complex features such as shapes.

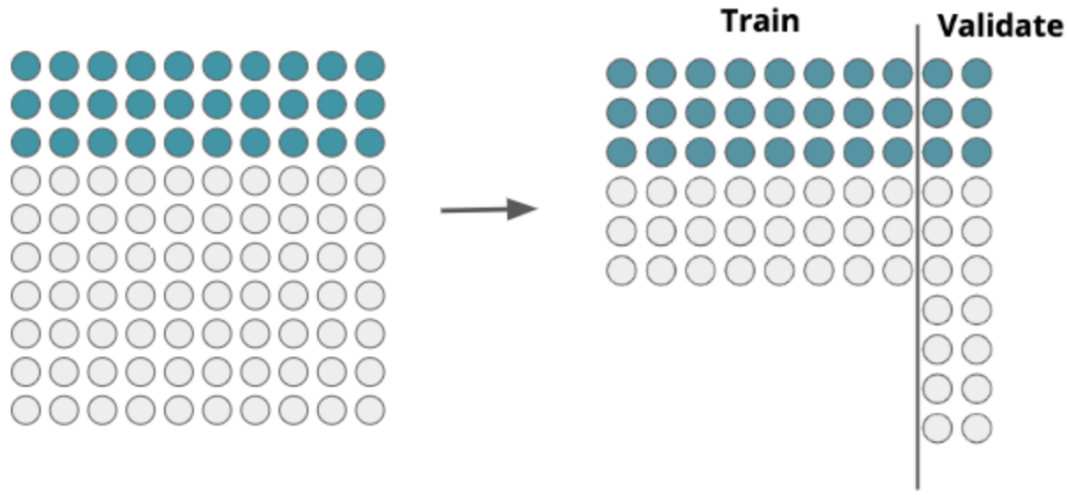
Note that the input image size must match the size of the *first* set of convolutional layers.

U-Net

U-Net is used for *segmentation* problems and it is more commonly used in 3D medical imaging. It's important to note that a limitation of 2D imaging is the inability to measure the *volume* of structures. 2D medical imaging only measures the area with respect to the angle of the image taken, which limits its utility in segmenting the whole area.

Split Dataset for Model Development

Splitting Your Data Example



You need to split your data into two sets before feeding it into the model.

- A training set: DL algorithm will use this data to learn the features that differentiate between your classes.
- A validation set: the algorithm will *never* use this set for learning. This is the set to determine if the algorithm is actually learning to discriminate between your classes.

The general rule of thumb is to split your data 80 in the training set and 20 in the validation set. The data should be split to maximize the *prevalence* of positive cases (i.e make sure 80% of your positive cases end up in the training set and 20% in the validation set).

We want to have a *balanced* training set so that the model has an equal number of cases in each class to learn. Even if one class is really rare in the wild. We want to have an *imbalanced* validation set to reflect the real-world situation.

For all other variables in your dataset such as age, sex, and race, the distribution should follow the *same* distribution as your original *full* dataset.

> 💡 **Note:** an image should NEVER be used for both training and validation.

Obtaining a Gold Standard

Gold standard

The *gold standard* for a particular type of data refers to the method that detects disease with the *highest* sensitivity and accuracy. Any new method that is developed can be compared to this to determine its performance. The gold standard is different for different diseases.

Ground truth

Often times, the gold standard is unattainable for an algorithm developer. So, you still need to establish the *ground truth* to compare your algorithm.

Ground truths can be created in many different ways. Typical sources of ground truth are

- Biopsy-based labeling. **Limitations:** difficult and expensive to obtain.
- NLP extraction. **Limitations:** may not be accurate.
- Expert (radiologist) labeling. **Limitations:** expensive and requires a lot of time to come up with labeling protocols.
- Labeling by another state-of-the-art algorithm. **Limitations:** may not be accurate.

Silver standard

The silver standard involves hiring *several* radiologists to each make their own diagnosis of an image. The final diagnosis is then determined by a *voting* system across all of the radiologists' labels for each image. Note, sometimes radiologists' experience levels are taken into account and votes are weighted by years of experience.

Image Pre-processing for Model Training

Image Pre-Processing

GOALS:

- Remove potential noise from your images (e.g. background extraction)
- Enforce some normalization across images (zero-mean, standardization)
- Enlarge your dataset (image augmentation)
- Resize for your CNN architecture's required input

Image Pre-Processing

GOALS:

- Remove potential noise from your images (e.g. background extraction)
- Enforce some normalization across images (zero-mean, standardization)
- Enlarge your dataset (image augmentation)
- Resize for your CNN architecture's required input


```
train_generator = train_datagen.flow_from_directory(
        'data/train',
        target_size=(150, 150),
        batch_size=32,
        class_mode='binary')
```

Intensity normalization

Intensity normalization is good practice and should always be done prior to using data for training. Making all of your intensity values fall within a small range that is close to zero helps the weights on our convolutional filters stay under control

There are two types of normalization that you can perform.

- zero-meaning: subtract that mean intensity value from every pixel.
- standardization: subtract the mean from each pixel and divide by the image's standard deviation.

Image augmentation

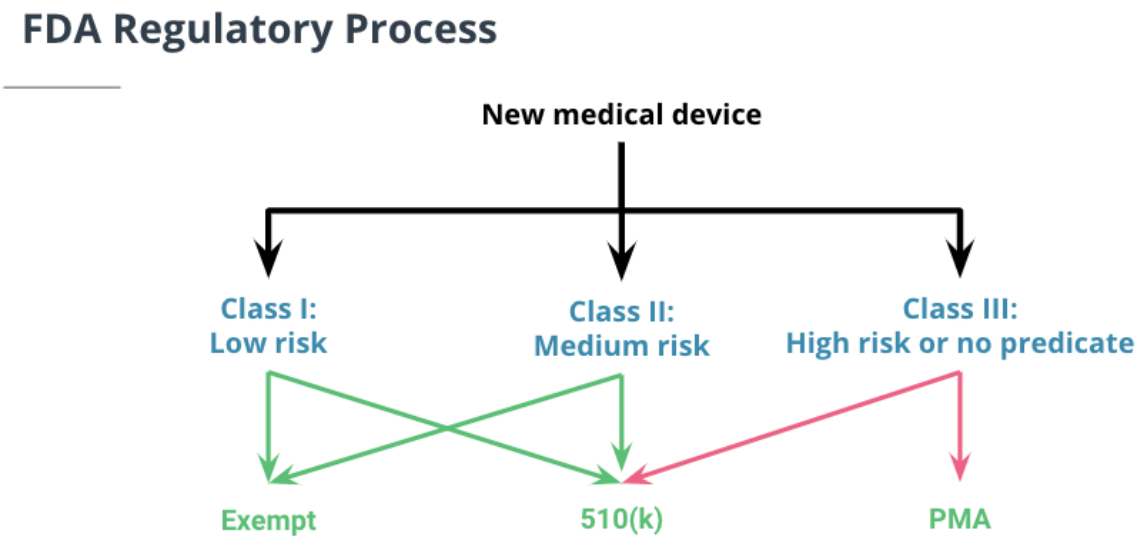
Image augmentation allows us to create different versions of the original data. Keras provides `ImageDataGenerator` package for image augmentation.

Note: not all image augmentation method is appropriate for medical imaging. A vertical flip should never be applied. And validation data should NEVER be augmented.

Image resize

CNNs have an input layer that specifies the size of the image they can process. Keras `flow_from_directory` have a `target_size` parameter to resize image.

Intended Use and Indications for Use



FDA Risk Catgories

Class I

> *"Not intended for the use in supporting or sustaining life or of substantial importance in preventing impairment to human health, and they may not present a potential unreasonable risk of illness or injury."*

💡 **47% of devices on the Market**

Class II

> *"devices for which general controls are insufficient to provide reasonable assurance of the safety and effectiveness of the device."*

💡 **Substantial equivalence to a predicate**

Class III

> *"Usually sustain or support life, are implanted or present a potential unreasonable risk of illness or injury."*

💡 **Only 10% of all FDA-regulated devices**

FDA Intended Use

FDA Intended Use

AI for 2D Medical Imaging Example

"...present a potential unreasonable risk of illness or injury."

Intended Use: for the identification of breast cancer from mammography.

Class III → Predicate exists → **501(k)**

Class III → No Predicate → **PMA**

AI for 2D Medical Imaging Example

"...present a potential unreasonable risk of illness or injury." ~~Class II ?~~

FDA assumes that the radiologist will rely on the software even though it's not labeled that way

Intended Use: for assisting the radiologist in the detection of breast abnormalities on mammogram.

Computer-Assisted Diagnosis (CADx)

Class III → Predicate exists → **501(k)**

Class III → No Predicate → **PMA**

Indications for use

Describe conditions for use

Intended Use: for assisting the radiologist in the detection of breast abnormalities on mammogram.

Indications for use:
Screening mammography studies
Women between the ages of 20-60 years old
With no prior history of breast cancer

Summary

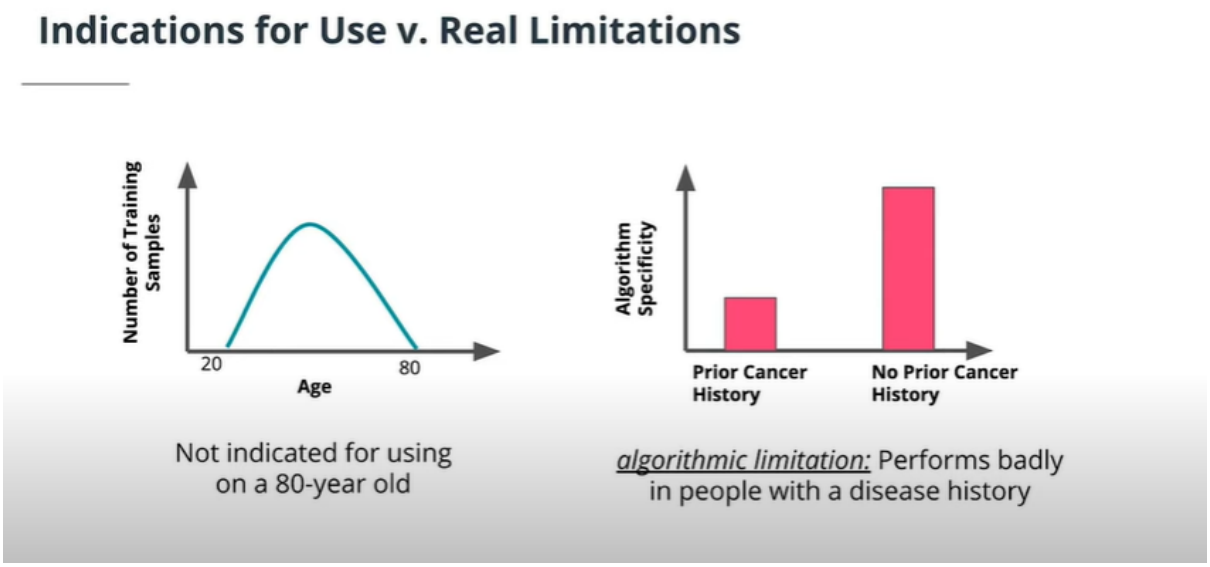
Intended use

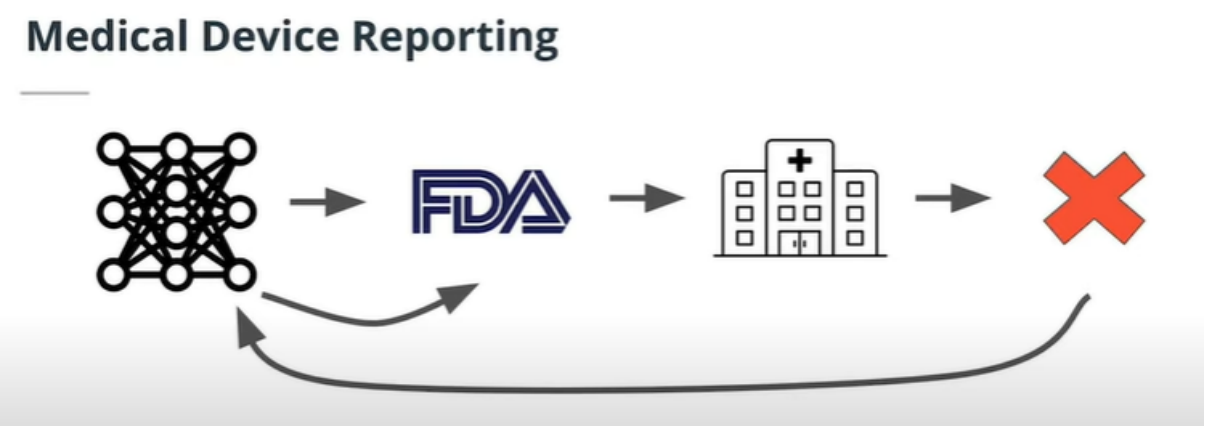
The FDA will require you to provide an intended use statement and an indication for use statement. The intended use statement tells the FDA exactly *what* your algorithm is used for. Not what it could be used for. And FDA will use this statement to define the risk and class of your algorithm.

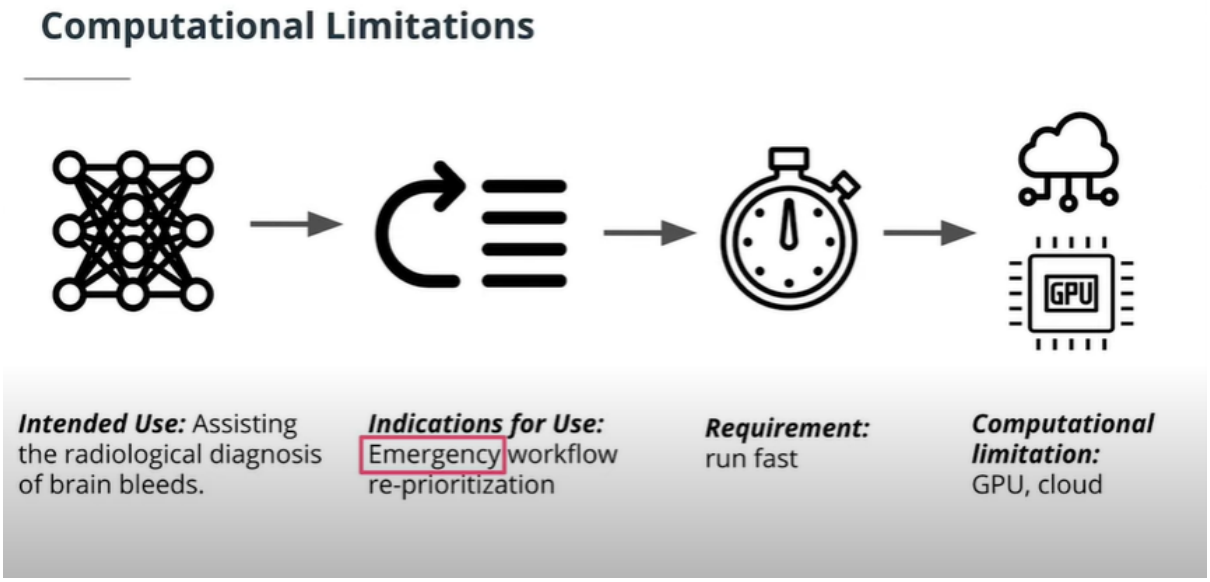
Indication for use

You can use the indications for use statement to make more *specific suggestions* about how your algorithm could be used. Indications for use statement describes precise situations and reasons *where and why* you would use this device.

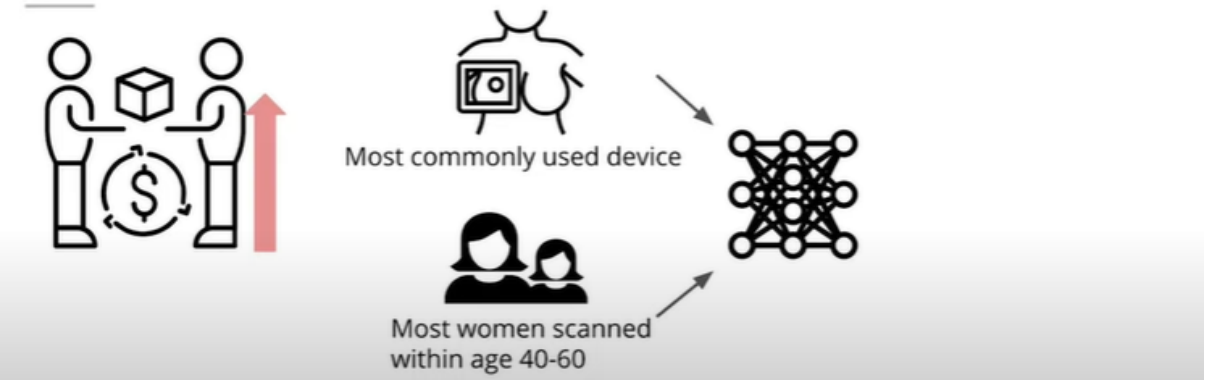
Algorithimic Limitations







Algorithmic Limitations' Imapct on Clinical Utility



Having a device that's very specific may actually be a good thing. Choosing a narrow scope when designing your devices is a good strategy, especially if you strategically choose something that maximises market imapct. For example, if you know that a particular scanner is the modt common type of

receiving screening mammograpy are between 40 and 60, then you still may end up with a incredibly useful and marketable algorithm if it is only labeled for use in women from 40-60 who are scanned by this particular scanner. You can always expand your functionality over time and work with the FDA to get new expansions approved.

Summary

Algorithm limitations

When the FDA talks about limitations, they want to know more about scenarios where your algorithm is not safe and effective to use. In other words, they want to know where our algorithm will *fail*.

Computational limitations

If your algorithm needs to work in an emergency workflow, you need to consider computational limitations and inform the FDA that the algorithm does not achieve fast performance in the absence of certain types of computational infrastructure. This would let your end consumers know if the device is right for them.

Medical device reporting

After your algorithm is cleared by the FDA and released, the FDA has a system called *Medical Device Reporting* to continuously monitor. Any time one of your end-users discovers a malfunction in your software, they report this back to you, the manufacturer, and you are required to report it back to the FDA. Depending on the severity of the malfunction, and whether or not it is life-threatening, the FDA will either completely recall your device or require you to update its labeling and explicitly state new limitations that have been encountered.

Translate Performance into Clinical Utility

Precision

Precision looks at the number of positive cases accurately identified by an algorithm divided by all of the cases identified as positive by the algorithm *no matter whether they are identified right or wrong*. This metric is also commonly referred to as the positive predictive value.

Precision and recall

A high precision test gives you more confidence that a positive test result is actually positive since a high precision test has low false positive. This metric, however, does not take false negatives into account. So a high precision test could still miss a lot of positive cases. Because of this, high-precision tests don't necessarily make for great stand-alone diagnostics but are beneficial when you want to *confirm* a suspected diagnosis.

When a high recall test returns a negative result, you can be confident that the result is truly negative since a high recall test has low false negatives. Recall does not take false positives into account though, so you may have high recall but are still labeling a lot of negative cases as positive. Because of this, high recall tests are good for things like screening studies, where you want to make sure someone *doesn't* have a disease or worklist prioritization where you want to make sure that people *without* the disease are being de-prioritized.

Optimizing one of these metrics usually comes at the expense of sacrificing the other.

Threshold

CNN models output a probability ranging from 0-1 that indicates how likely the image belongs to a class. We will need a cut-off value called threshold to assist in making the decision if the probability is high enough to belong to one class. Recall and precision vary when a different threshold is chosen.

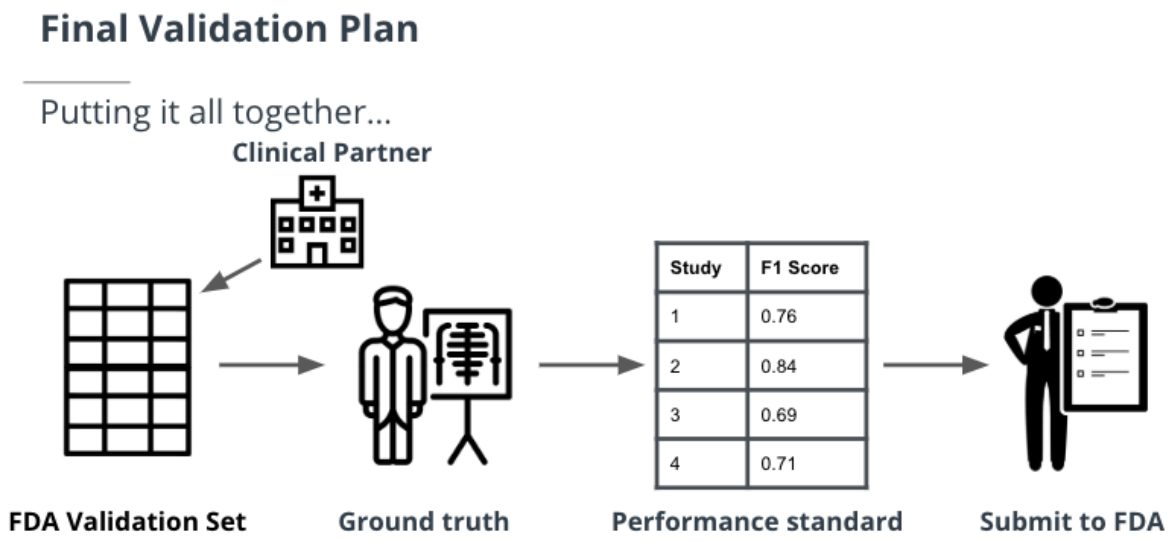
Precision-recall curve

Precision-recall curve plots recall in the x-axis and precision in the y-axis. Each point along the curve represents precision and recall under a different threshold value.

F1 score

For binary classification problems, the F1 score combines both precision and recall. F1 score allows us to better measure a test's accuracy when there are class *imbalances*. Mathematically, it is the harmonic mean of precision and recall.

Designing an FDA Validation Plan



FDA validation set

You'll need to perform a standalone clinical assessment of your tool that uses an *FDA validation set* from a real-world *clinical setting* to prove to the FDA that your algorithm works. You will run this FDA validation set through your algorithm just ONCE.

You'll need to identify a clinical partner who you can work with to gather the "BEST" data for your validation plan. This partner will collect data from a real-world clinical setting that you describe so that you can then see how your algorithm performs under these specifications.

Collect the FDA validation set

You need to identify a clinical partner to gather the FDA validation set. First, you need to describe who you want the data from. Second, you need to specify what types of images you're looking for.

Establish the ground truth

You need to gather the ground truth that can be used to compare the model output tested on the FDA validation set. The choice of your ground truth method ties back to your *intended use* statement. Depending on the intended use of the algorithm, the ground truth can be very different.

Performance standard

For your validation plan, you need evidence to support your reasoning. As a result, you need a performance standard. This step usually involves a lot of

literature searching.

Depending on the use case for your algorithm, part of your validation plan may need to include assessing *how fast* your algorithm can read a study.

Designing an FDA Validation Plan Exercise

Algorithm A:

Intended Use: Assisting a radiologist with *classifying breast density*

Indications for Use: indicated for use in screening mammography studies in women of ages 40-80. *Not* indicated for use in women with artificial implants.

Hint: Breast density falls into four categories: A, B, C, D and can only be determined from an image. It is notoriously difficult to determine breast density on a mammography study, but that's because there is no "correct" answer. It's a sliding scale that radiologists tend to disagree on when a patient is right on the border between two density levels. Also, in the real world, breast densities A & D are both about 10% prevalent, while breast densities B & C are both about 40% prevalent.

Algorithm B:

Intended Use: Assisting a radiologist with identifying *breast abnormalities*.

Indications for Use: indicated for use in screening mammography studies in women for ages 40-80. *Not* indicated for women with a prior history of breast cancer.

Hint: Radiologists are *really* good at detecting a wide range of abnormalities on screening mammograms, and a radiologists' read is considered the gold standard for determining 'normal' vs. 'abnormal' for an imaging study.

In the free response section, compare and contrast how you would do the following for your FDA validation plan for each of the two algorithms:

- Define the clinical population needed for the validation data set that you obtain from your clinical partner
- Choose the method of obtaining a ground truth

Solutions

For both algorithms, I would want to collect a validation set that was made up of screening mammography studies only for women between the ages of 40 and 80.

To validate algorithm A, however, I would want to make sure that there were no implants, and I would also want to make sure that the distribution of breast densities A, B, C, and D was reflective of the distribution of those densities that are seen in the real world.

To validate algorithm B, I would want to make sure that my validation data set did not contain any women who had a prior history of breast cancer.

The silver standard approach of using several radiologists would be more optimal for Algorithm A because I gave you the hint that it's really hard for radiologists to agree on breast density labels.

For Algorithm B, a single radiologist's labels would probably suffice, because I gave you the hint that they're really good at labeling 'normal' v. 'abnormal.'