

NCI Virtual Workshop on Medical Imaging De-Identification (MIDI)

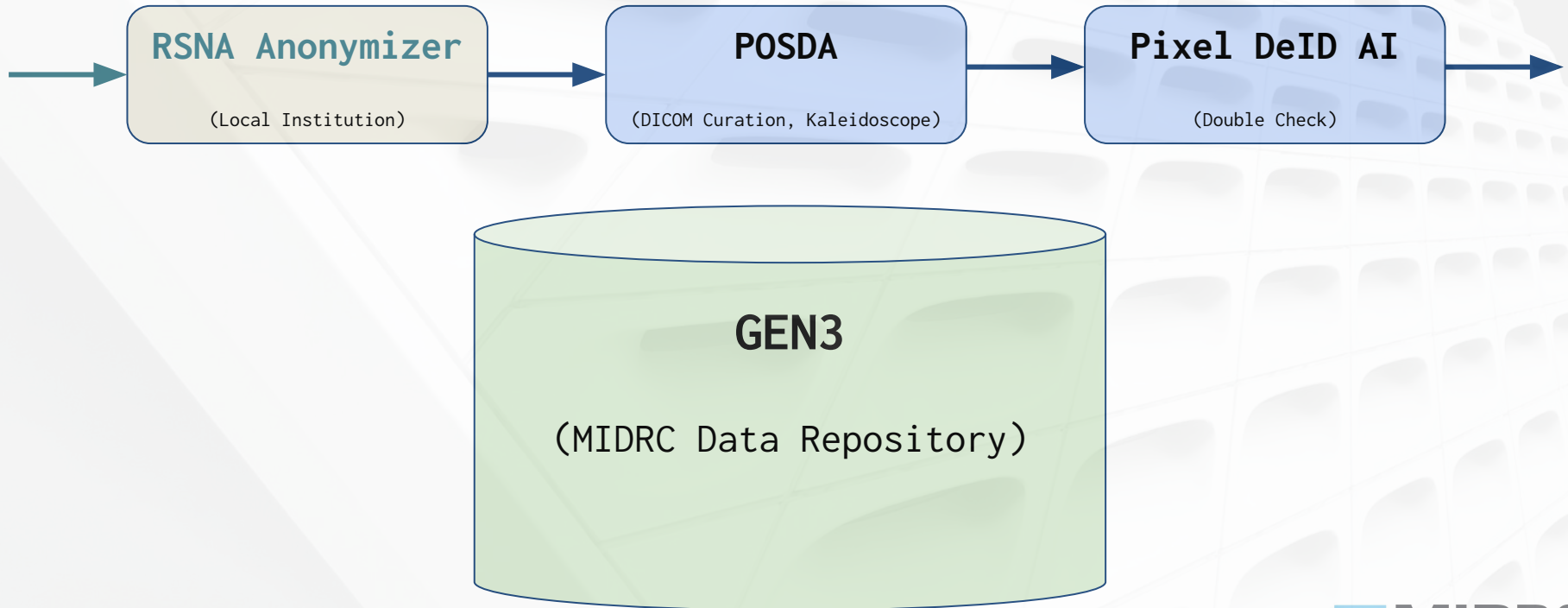
Pixel De-Identification Using AI

May 23, 2023 12:00pm ET

George Shih, MD MS FACR
Professor and Vice-Chair for Informatics
Weill Cornell Radiology

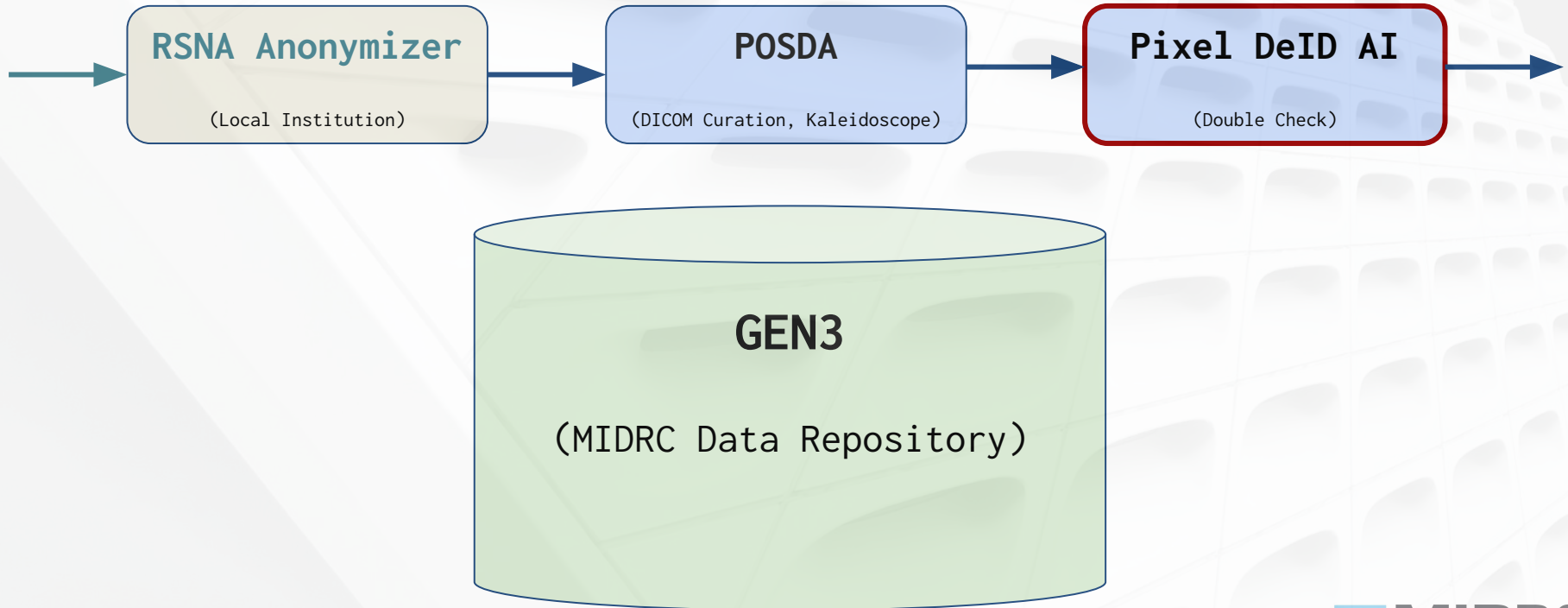
RSNA MIDRC

Data Workflow



RSNA MIDRC

Data Workflow



Patient Name: [REDACTED] Exam no: 1744
Accession Number: [REDACTED]
Patient ID: [REDACTED] Discovery CT750 HD
Exam Description: CT HALS/THORAX/ABDOMEN

Dose Report

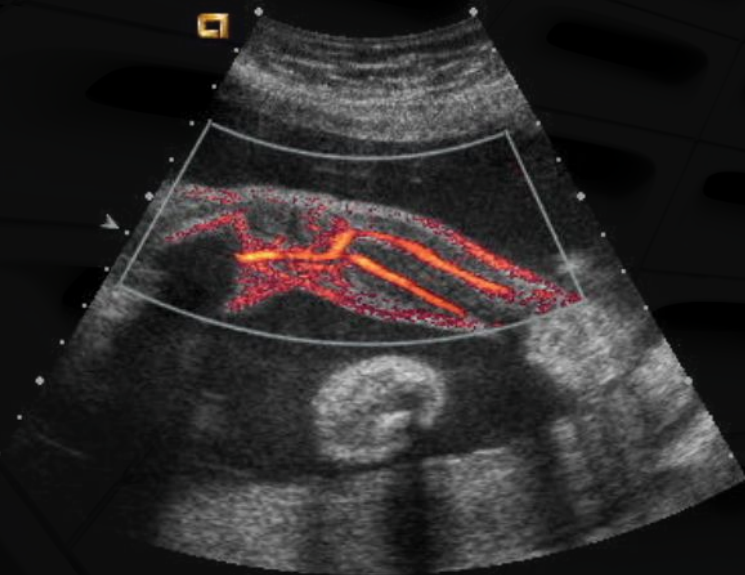
Series	Type	Scan Range (mm)	CTDIvol (mGy)	DLP (mGy-cm)	Phantom cm
1	Scout	-	-	-	-
2	Helical	S15.750-I650.250	5.10	373.00	Body 32
5	Helical	S188.000-I105.000	5.10	182.72	Body 32
Total Exam DLP:				555.72	



SEQUOIA



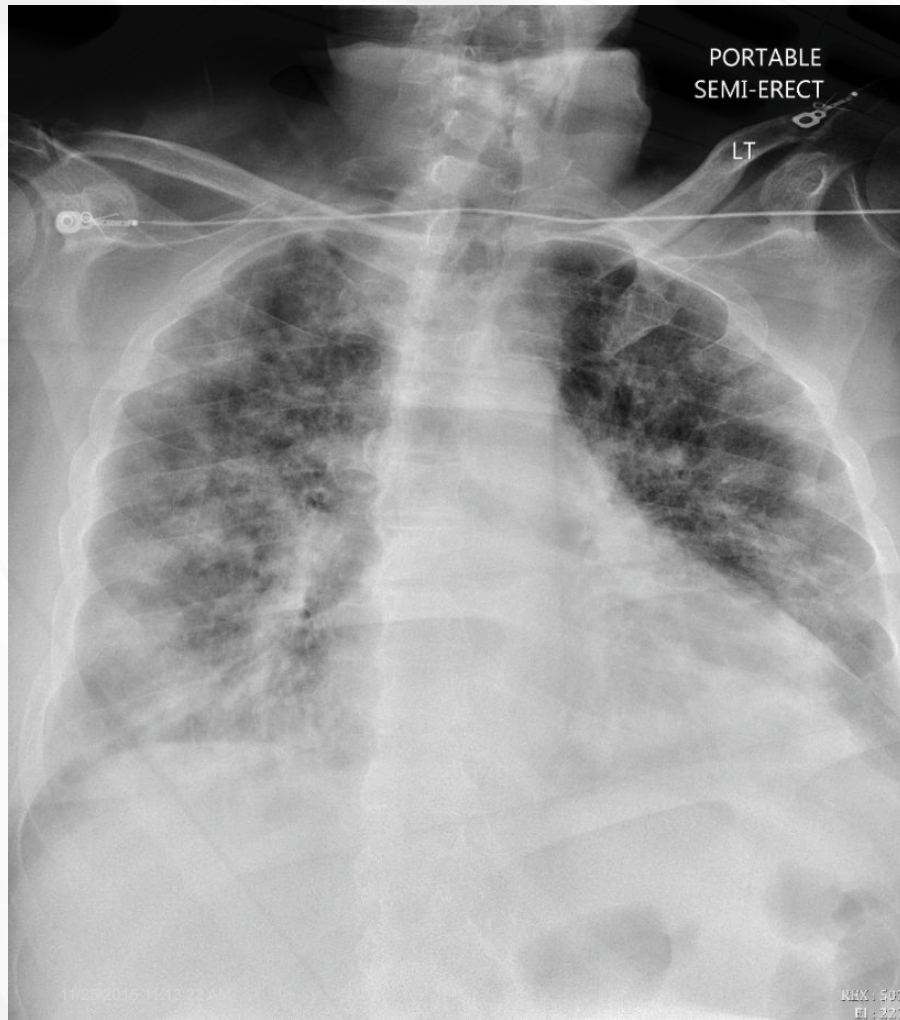
.089

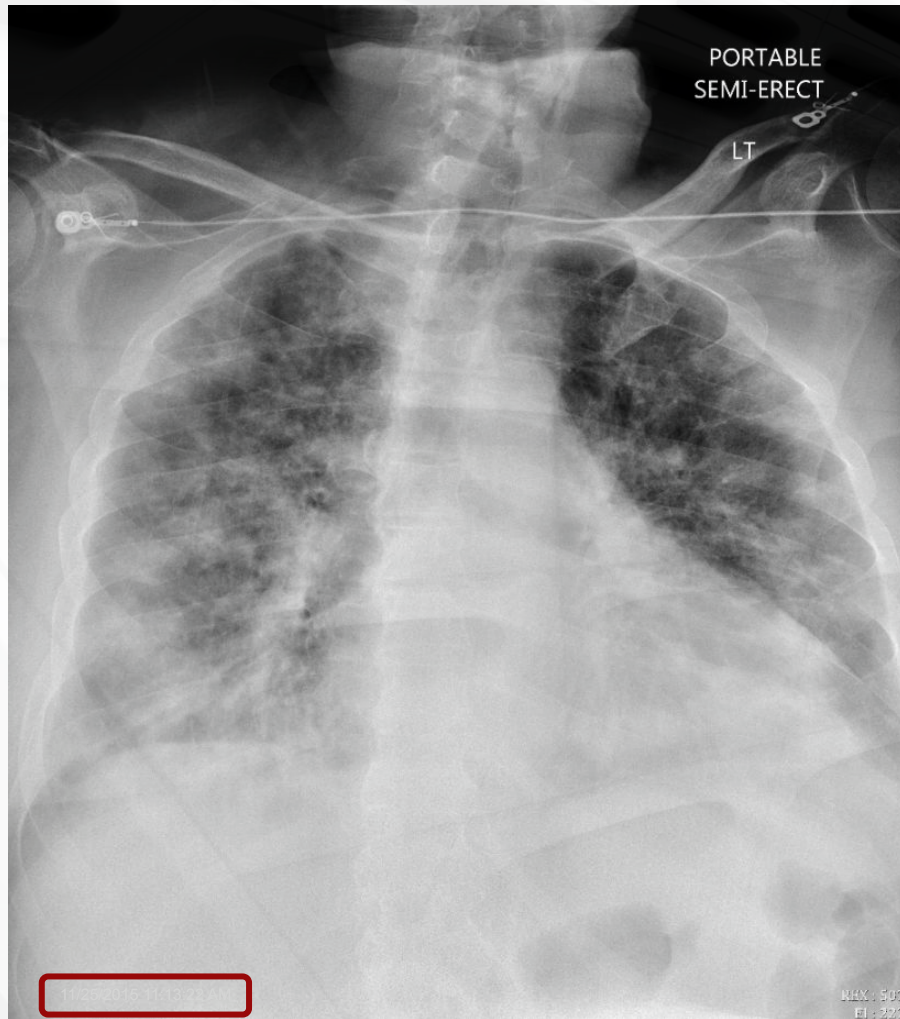


6C2 4Hz
H5.0MHz 30mm
OB
General

S2/ 0/ 3/E:1+1
1/1 CD:5MHz
CD Gain = 43
CDE 15dB

Store in progress



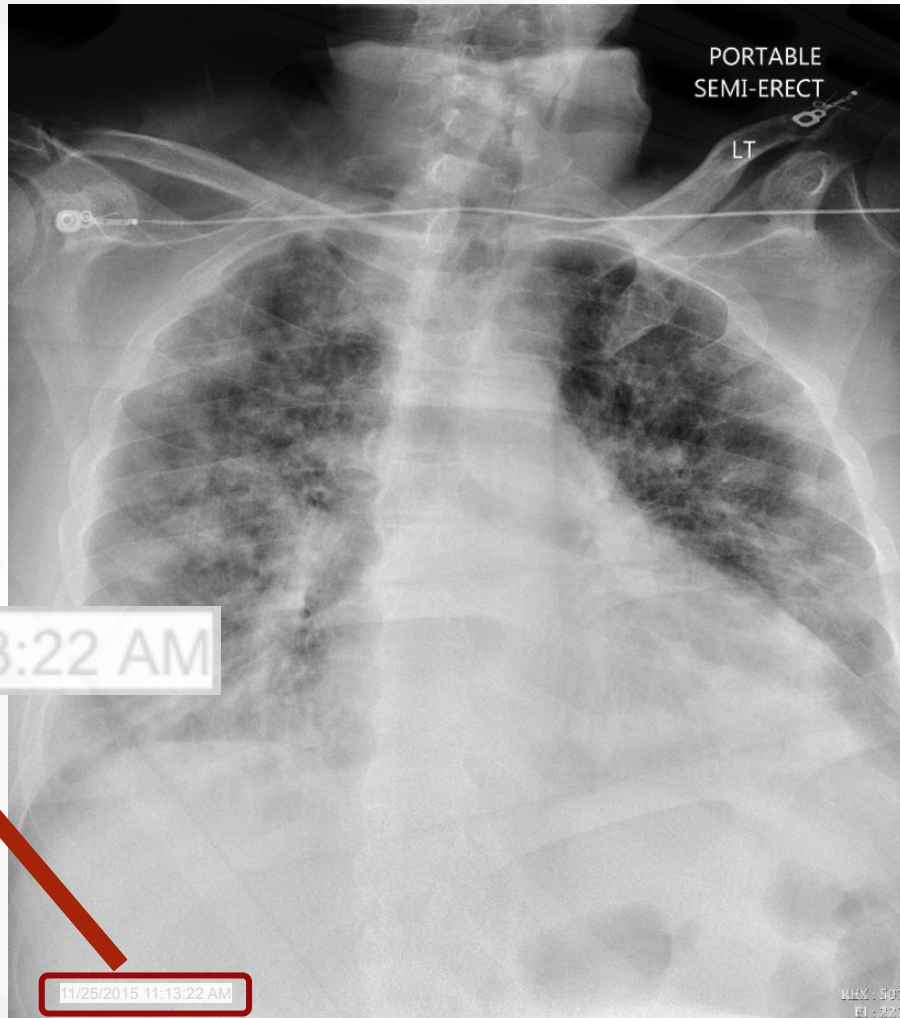


PORTABLE
SEMI-ERECT

LT

1/29/2016 11:13:21 AM

MAX: 500
EI: 220



11/25/2015 11:13:22 AM

11/25/2015 11:13:22 AM

MAX: 500
EI: 200

Pixel DeID AI



Leveraging AI

- Manual de-identification is time consuming, requires significant resources and prone to human error.
- Many tools require specific configuration files or manual selection that may fail to capture PHI in all scenarios
- CNN/transformer based models scalable, flexible and adaptable across different modalities.
- Large language models (eg. GPT4) may help improve and make post processing of outputs more robust.

Pixel DeID Workflow

- Three step process



DAVIDSON DOUGLAS [M] 01.09.2012
DOB: 06.16.1976

DAVIDSON DOUGLAS M 01092012
DOB 06161976



PHI Detection and Classification

- Deep learning model that purely looks at pixel data and detects text on images. Classifies outputs into three categories
 - **Text** (potential PHI)
 - **Digits** (text with any digits, potential PHI)
 - **Markers** (radiological markers, may not want to redact)

Optical Character Recognition (OCR) + GPT-4

- Detected bounding boxes are passed through a robust OCR model to extract accurate textual information.
- Currently a deep learning + NER approach classifies predictions into Text, Digits or Markers. Can also utilize GPT-4 for zero shot or one shot classification -

Playground

Load a preset...

Save

View code

Share

...

SYSTEM

You are an expert in de-identifying medical images.

You will be given comma separated instances of text that were passed through an OCR engine. You need to fix the outputs in case they are wrong and classify the text into one of three categories -

1. Digits - Any text that contains numbers that could be potential PHI like phone number, DOB etc.
2. Text - Any text that could be potential PHI
3. Markers - Non PHI data that includes radiological markers that may be helpful like L, R, Portable etc.

Individually fix the text and give the prediction for each instance separated by a colon in new lines.

USER

I, 21, dob06161976, semiupright, David Douglas, rcc

ASSISTANT

L: Markers
21: Digits
DOB 06-16-1976: Digits
Semi-upright: Markers
David Douglas: Text
RCC: Markers

+ Add message

Mode

Chat Beta

Model

gpt-4

Temperature

0

Maximum length

256

Top P

1

Frequency penalty

0

Project Labels Navigation

Workspace

Jupyter

Users

Meet

Chat

Progress

Export

Info

Settings

Community

Docs

Labels

Navigation

Dataset: PHI Dataset

Data View: Default Hierarchical

Columns: 1

Filter Label Group

Text Bounding Box

Digits Bounding Box

Markers Bounding Box

Exam 1/16

Series 2/2

Image 1/1

Notes

Models

De-ID

NEW LABEL

CLONE LABEL GROUP

LABEL CONTROLS

NEW MODEL

CLONE MODEL

MODEL REGISTRY

COLLAPSE ALL

VIEW MODEL PERFORMANCE

WARMUP ALL MODELS

UPLOAD DATA & INFERENCE

INACTIVE

Pixel De-ID

Model for detecting burnt in patient health information and markers on DICOM images with an emphasis on detecting dates/digits.

OWNER: M_KlnqyN

SCOPE: Image

CREATED: 4 months ago

UPDATED: 4 months ago

PRIVACY: [lock icon]

AUTORUN: [off icon]

LABELS WITH CLASS INDEX

[0] Text [1] Digits [2] Markers

VERSIONS

#	ID	Docker Image:Tag	Created	Build Status	Hardware
✓ 3	660	staging/model-244:v3	4 months ago	Success	GPU

1-1 of 1

NEW MODEL VERSION

TASKS

ID	Created	Type	Status	Progress
1306	4 months ago	Inference	Succeeded	100%

1-1 of 1

Run model version 3 on the active:

Image for: Inference

RUN INFERENCE

SAVED

DAVIDSON DOUGLAS [M] 01.09.2012

DOB: 06.16.1976

Semi-Upright

Portable

L

MODEL OUTPUT [ID: 0_11420e]

Created 5 months ago

Model: Pixel De-ID [v3]

Task: ID: 1306

Class Name: Digits

Class Probability: 0.7078903416211809

Explanations: cmicsoadddugaasm11122222

DAVIDSON DOUGLAS [M] 01.09.2012

Text Markers Markers Digits Markers Digits

Redaction

- Pixel redaction replaces text with black box.
- Can redact all predictions or choose a subset to redact (human in the loop)

The screenshot shows the 'Create Dataset De-ID Task' interface. On the left is a navigation sidebar with buttons for Exam (1/19), Series (1/2), Image (1/1), Notes, Models, and De-ID. The main panel is titled 'Create Dataset De-ID Task' and includes a 'Dock Panels' toggle, 'Dataset' (PHI Dataset [modi...]), and 'Data View' (Default Hierarchical) dropdowns. Three task options are listed: 'DICOM Pixel Redaction' (checked), 'Apply Automatic Inference' (unchecked), and 'DICOM Tags Modification' (checked). The 'PIXEL REDACTION' configuration panel is highlighted with a red border and contains a 'Select Labels' dropdown menu with 'Text' and 'Digits' selected. Below this is the 'TAGS MODIFICATION' section with '+ ADD ROW', 'EXPORT CSV', and 'CLEAR' buttons, and a checkbox for 'Remove Private DICOM Tags'. A table lists tags to be modified, and a 'CREATE TASK' button is at the bottom.

PIXEL REDACTION X CLEAR

Select Labels

Text Digits

TAGS MODIFICATION + ADD ROW EXPORT CSV X CLEAR

Remove Private DICOM Tags LOAD PRESET

Rows per page: 5 1-5 of 147

Action	DICOM Keyword i	Value
REMOVE	AcquisitionContextSequence	🗑
JITTER	AcquisitionDate	🗑
JITTER	AcquisitionDateTime	🗑
REMOVE	ActualHumanPerformersSequence	🗑
REMOVE	AdmissionID	🗑

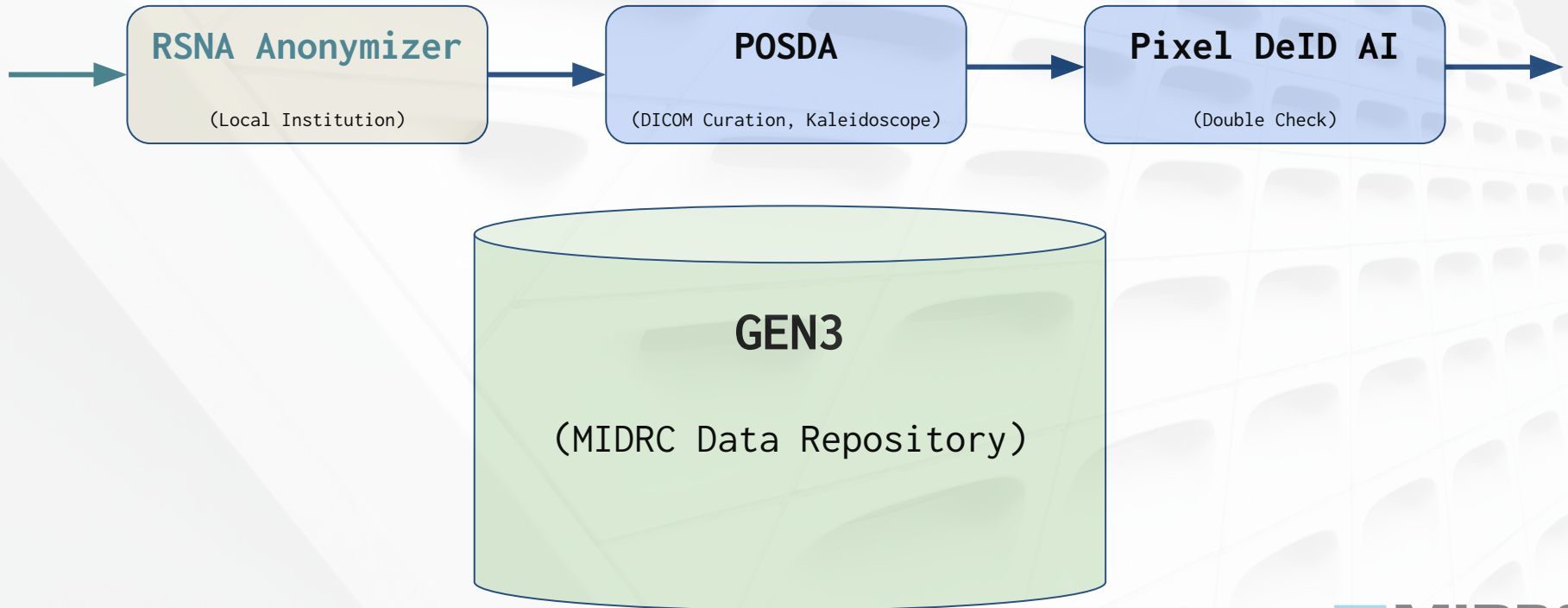
➡ CREATE TASK



DeID Workflow

RSNA MIDRC

Data Workflow



1. RSNA Anonymizer For DICOM Tags

- The Admin tab of Anonymizer provides access to the script used in de-identification.
- You can also generate a table (in MS-Excel) that shows the mapping of original to pseudonymous values for patient name and ID and study date, as well as the integer representation of the date offset.
- To generate the mapping table, click the Admin tab, then Index, then the List and Save buttons.
- Keep the table secure, as you would any document containing PHI.

The screenshot displays the RSNA Anonymizer software interface. The top menu bar includes 'Welcome', 'Q/R SCU', 'Storage SCP', 'Directory', 'Export', 'Admin', and 'Help'. The 'Admin' tab is selected and circled in red. Below the menu bar, there are sub-tabs: 'Viewer', 'Elements', 'Filter', 'Script', 'Index', and 'Log'. The 'Index' tab is also circled in red. The main area shows a 'Patient Index List' with the following data:

ANON-PatientName	ANON-PatientID	PHI-PatientName	PHI-PatientID
316093-000001	316093-000001	RSNA29bfbb3	RSNA29bfbb3

Below the list, an Excel spreadsheet is open, showing a mapping table. The spreadsheet has columns A through J and rows 1 through 8. The data in row 2 is as follows:

	A	B	C	D	E	F	G	H	I	J
1	ANON-PatientName	ANON-PatientID	PHI-PatientName	PHI-PatientID	DateOffset	ANON-StudyDate	PHI-StudyDate	ANON-Accession	PHI-Accession	
2	316093-000001	316093-000001	RSNA29bfbb3	RSNA29bfbb3	492	20141016	20160220	1	RSNA29bff95	
3										
4										
5										
6										
7										
8										

At the bottom of the Excel window, the 'List' and 'Save' buttons are circled in red.

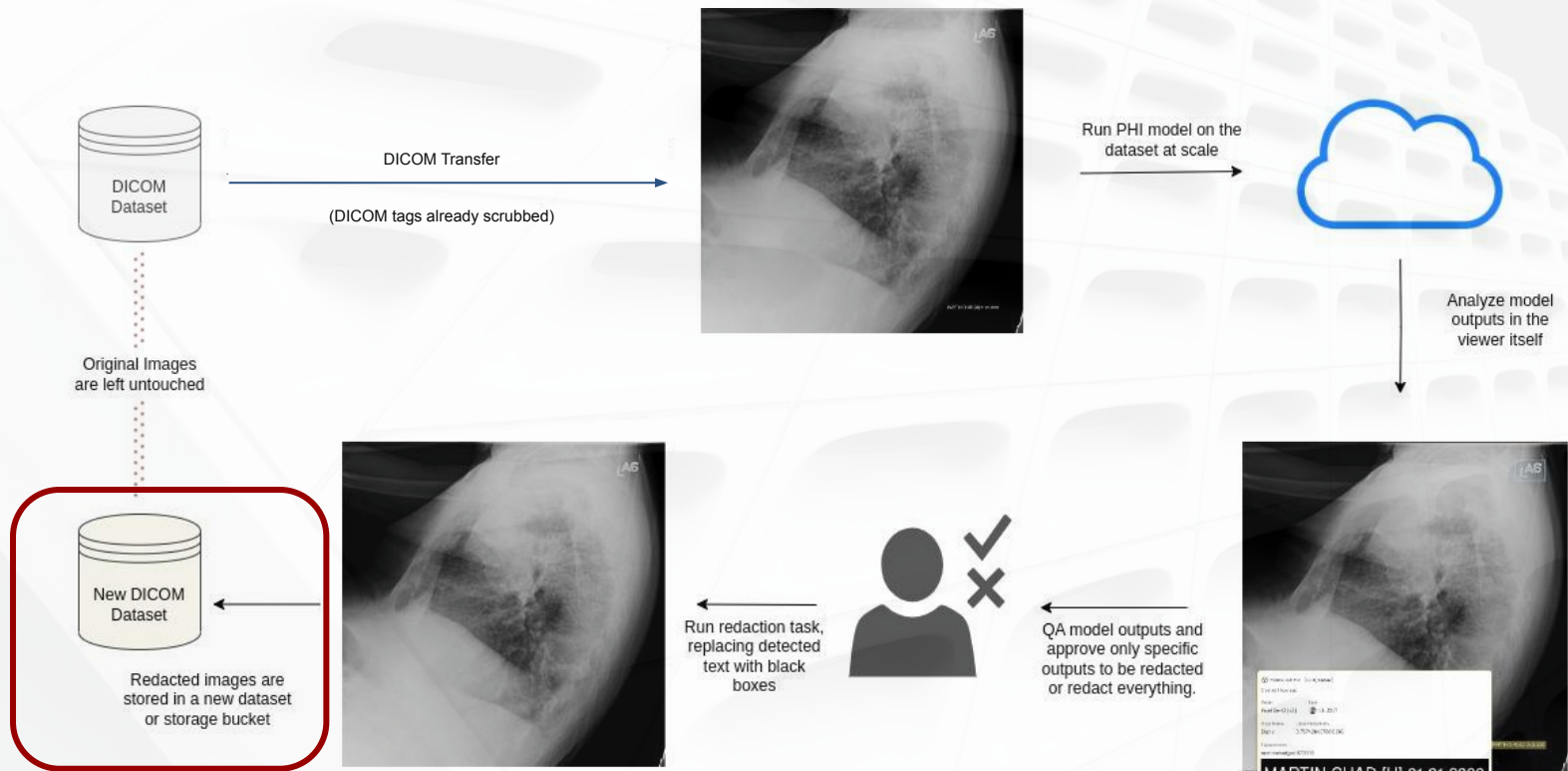
2. Human Visual Inspection Using Posda Tools

- Once Kaleidoscope opens you will have the option to review all the images, mark images as Good or Bad, etc. and to perform further review with the Quince viewer.
- See the Using Kaleidoscope and Quince section in the POSDA user Guide for more information on the viewers.

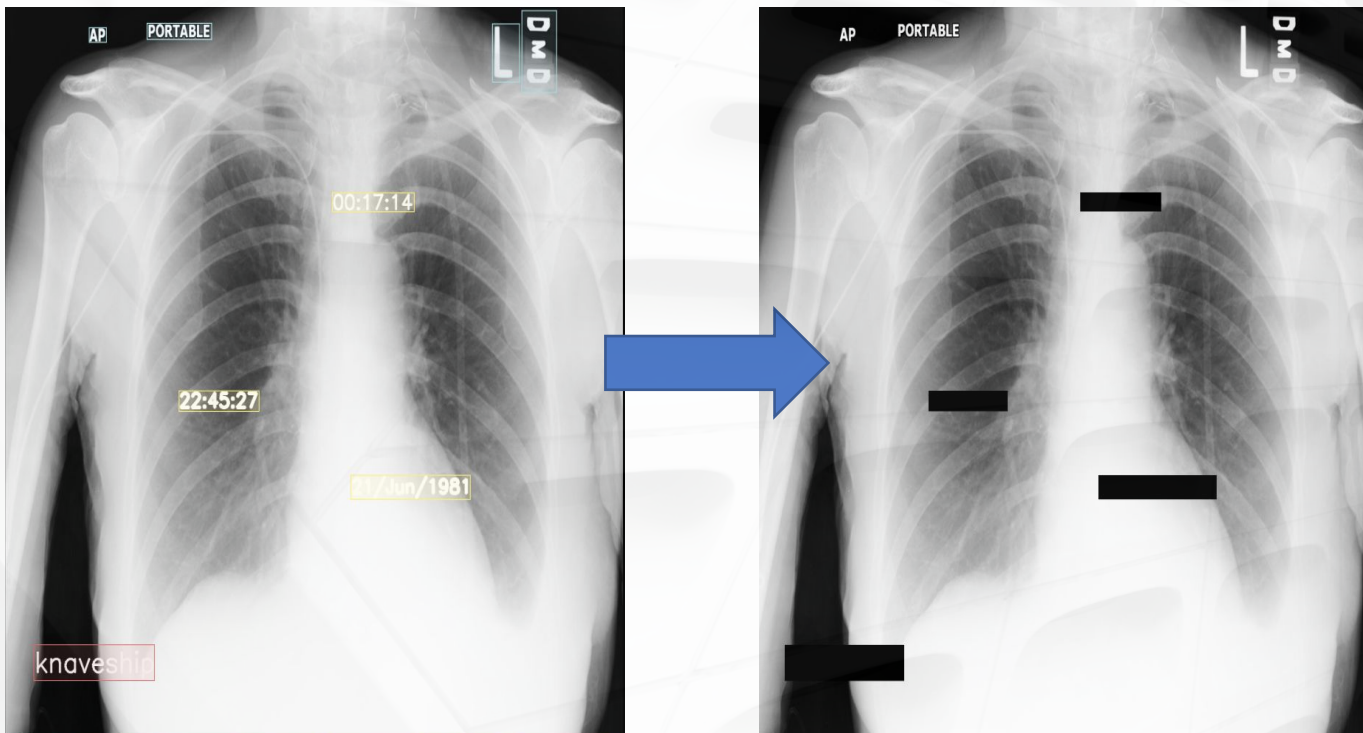
The image shows two overlapping software interfaces. The background interface is 'Kaleidoscope', which displays three X-ray images of a human torso. At the top, there are navigation buttons: 'Go Back' and 'Skip Forward' (circled in red), and a row of status buttons: 'Good' (green), 'Bad' (red), 'Blank' (orange), 'Scout' (yellow), and 'Other' (grey). Below the images, there is a table with patient information and a button labeled 'Open in Quince' (circled in red). The foreground interface is 'Quince: DICOM Viewer', which shows a single X-ray image with a zoom slider and playback controls. At the bottom, there are buttons for 'Presets', 'Reset Window/Level', 'Reset Zoom', 'ROI', 'Details', and 'Dump'. A second 'Open in Quince' button is circled in red at the bottom right of the Quince viewer.

IEC	Open in Quince
Images in IEC	2
Patient ID	Pat_796
Body part examined	KIDNEY
Series Instance UID	1.3.6.1.4.1.14519.5.2.1.888.4001.155749122736545709243854288452
Path	/home/posda/cache/k-storage/9b/47/44/9b4744e07260f5851cf92ed85a2a9e8f

3. Pixel DeID AI Workflow

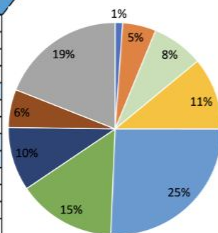
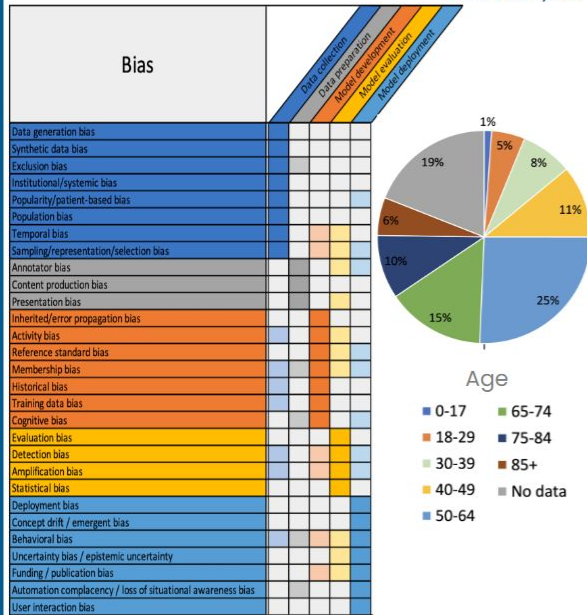


3. Pixel DeID AI Workflow



Bias & Diversity Working Group

A diverse data collection and curation strategy, as well as the mitigation of bias in data analysis within the MIDRC commons, are critically important to yield ethical AI algorithms that produce trustworthy results for all groups. MIDRC strives to mitigate bias in its study population, data collection, curation and analysis.



Grand Challenges Working

- Performs oversight of all MIDRC Grand Challenge processes, protocols, and guidelines
- Maintains MIDRC Portfolio of Grand Challenges
- Planning several upcoming Challenges using pre-published MIDRC data with expert annotations, including:



- COVID-19 severity
- Patient outcome prediction using both clinical and imaging data
- Longitudinal assessments, disease progression
- Risk assessment for long-term sequelae (e.g., neuro, cardiovascular)
- COVID segmentation on CT data

MIDRC COVIDx CHALLENGE



A COVID classification Grand Challenge on pre-published portable chest radiographs from MIDRC

Top-ranked finishers will be acknowledged during our session at the Innovation Theater, Booth 3316, South Hall Tuesday, November 29, 4:00 PM - 5:00 PM

Cash awards generously sponsored by the International Society for Photonics and Optics (SPIE)

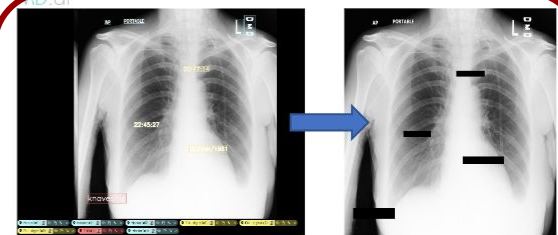
Annotations Working Group

Crowd-Sourced Annotations

- To create standard methods for labeling and annotation of COVID-19 radiographic and CT images.
- To develop best-practices for image labeling by crowds of imaging experts, including:
 - Recruitment and training of experts
 - Image label adjudication
 - Methods to assess accuracy of experts
 - Data cleaning methods for labels.
- To deploy these tools and processes to conduct 3 data science challenges to develop computer vision algorithms for COVID-19.

Helper AI & the Annotation Process:

- High quality, well-curated annotations are an essential supplement to the MIDRC effort and AI research
- Annotations can be created by human experts or automatically created by tools like AI models.



- MD.ai's recent application of Helper AI: Using an AI model to help identify pixel data in regions containing PHI. The burned-in PHI is then redacted using the AI model's annotations.

Annotation Members:

RSNA - Adam Flanders, Bhavik Patel, Carol Wu, Chris Carr, George Shih, Maryam Vazirabad, Jason Sho, Thomas OSullivan
 ACR - Brian Bialecki AAPM - Andrey Federov, Paul Kinahan, Sam Armato
 Gen3 - Bob Grossman, UChicago-MIDRC Central - Nick Gruszauskas
 MD.ai team: Chinmay Singhal, Zhihao Wang

BDWG Members:

AAPM - Weijie Chen, Karen Drukker, Kyle Myers, Berkman Sahiner, Emily Townley ACR - Jayashree Kalpathy-Cramer, Judy Wawira-Gichoya NIH - Rui de Sá RSNA - Sanmi Koyejo, Zi Jill Zhang UChicago-MIDRC Central - Maryellen Giger, Nick Gruszauskas, Heather Whitney



MIDRC-GCWG Members:

AAPM - Sam Armato, Karen Drukker, Lubomir Hadjijski, Emily Townley ACR - Jayashree Kalpathy-Cramer, Chris Tremel NIH - Rui de Sá RSNA - Robyn Ball, Adam Flanders, Tim Stearns, Carol Wu UChicago-MIDRC Central - Maryellen Giger, Ravi Madduri

