AI and De-Identification for Medical Imaging within CRDC

Granger Sutton, Ph.D.

October 17, 2024

Introduction

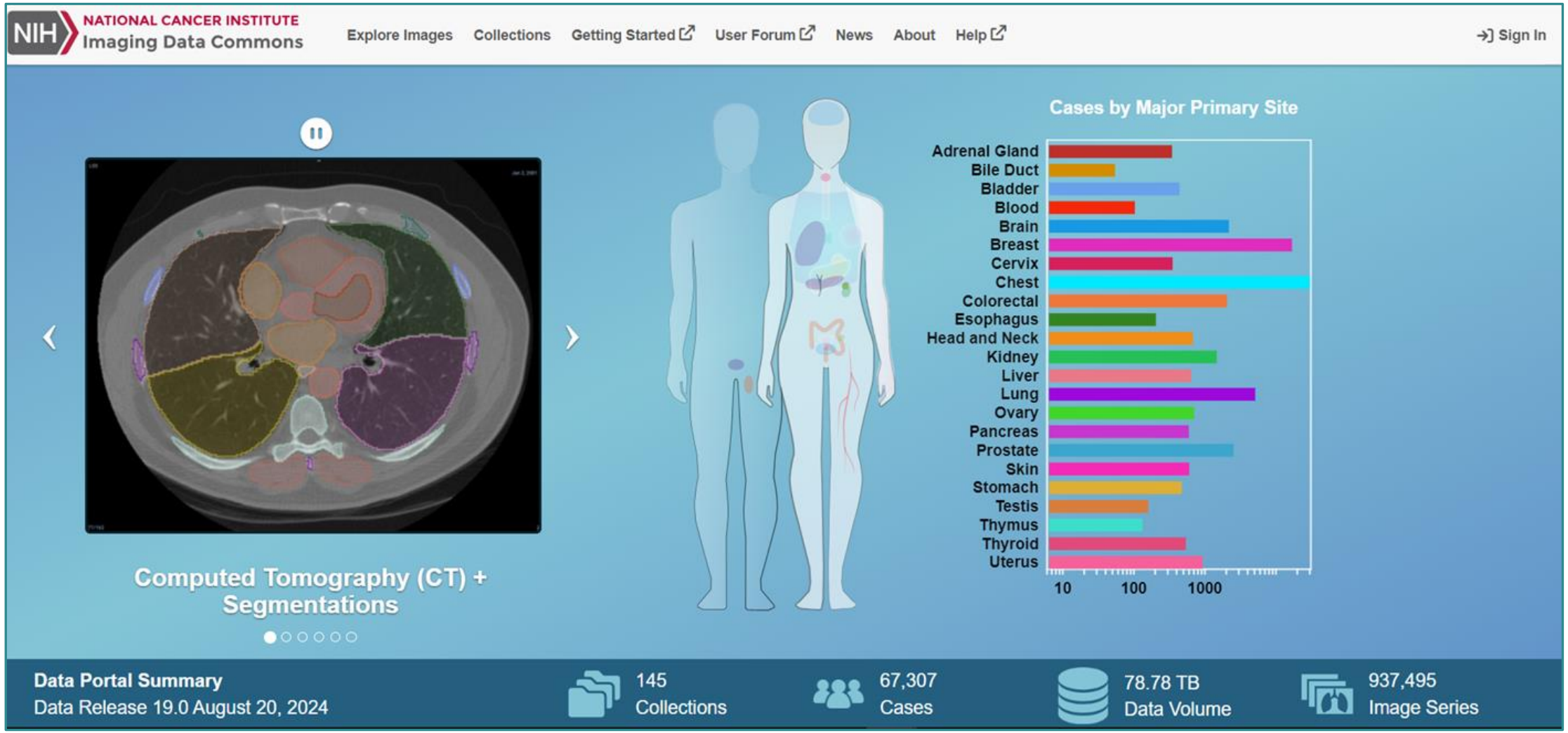
- AI -> IDC -> AI
- Two AI Projects: MIDI and MHub
  - MIDI: Medical Image De-Identification
  - MHub: Model Hub - AI model platform
- Leverage DICOM standard

Leveraging DICOM imaging file standard

- Digital Imaging and Communications in Medicine (DICOM)
- Standard interface for AI Tool Development
- Data Harmonization and Interoperability
- Supports AI-generated Features overlaid on images

The International Organization for Standardization recognizes DICOM® as the ISO 12052 standard

Imaging Data Commons (IDC)



https://imaging.datacommons.cancer.gov

Imaging Data Commons (IDC)

- Data exploration, visualization, cohort building
- Free download and accessible in the cloud
- Tutorials and Colab notebook examples for applying AI models to IDC data
- >78 TB data from numerous studies:
  - Radiology
  - Digital pathology
  - Fluorescence
  - AI-derived features
  - Clinical data
  - DICOM image file standard

Enhance IDC Data: AI addresses Needs

- Identifiable patient data in images
  - **Need:** a tool to automate de-identification
  - **Solution:** Medical Image De-Identification (MIDI) tool
- AI can enhance the use of imaging datasets
  - **Need:** an AI model platform to build, store, and provide reproducible image analysis
  - **Solution:** Model Hub (MHub) platform

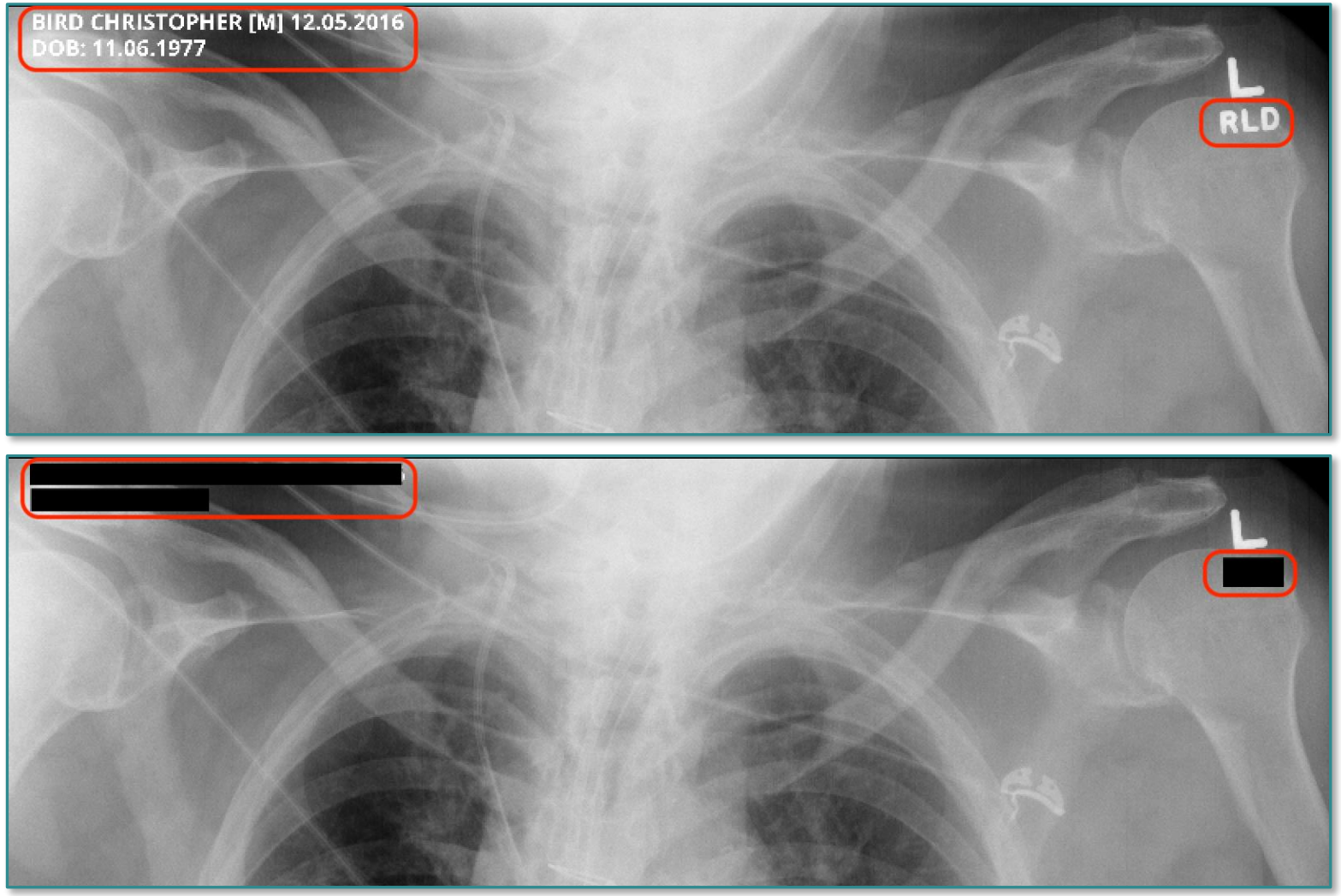
Medical Imaging De-Identification (MIDI)

- To ensure patient privacy, medical images must be de-identified before sharing
- Compliance with legal requirements:
  - Health Insurance Portability and Accountability Act (HIPAA)
  - General Data Protection Regulation (GDPR)
- DICOM simplifies de-identification
- DICOM has explicit rules for de-identifying images

Medical Imaging De-Identification (MIDI)

- Identifiable data in IDC DICOM files are currently manually removed at TCIA
- Save time and money with a comparable automated de-identification tool
  - External: Images can be de-identified at low cost
  - Internal: CRDC can validate de-identification
- Not yet available to the public
- Final testing: real data - manual vs automated

PHI burned into DICOM Images



- **Top:** before de-identification
- **Bottom:** after de-identification
- **Note:** DICOM examples all presented with synthetic PII/PHI

PHI in DICOM Header: Before MIDI

DICOM Tags

Search...

(Group,Ele...	TAG Description	Value
(0002,0000)	FileMetaInformationGroupLength	194
(0002,0001)	FileMetaInformationVersion	
(0002,0002)	MediaStorageSOPClassUID	1.2.840.10008.5.1.4.1.1.13.1.3
(0002,0003)	MediaStorageSOPInstanceUID	2.3.761.0.1.2402815.6.794.9587984116011978201
(0002,0010)	TransferSyntaxUID	1.2.840.10008.1.2.4.90
(0002,0012)	ImplementationClassUID	1.3.6.1.4.1.22213.1.143
(0002,0013)	ImplementationVersionName	0.5
(0002,0016)	SourceApplicationEntityTitle	POSDA
(0008,0005)	SpecificCharacterSet	ISO_IR 100
(0008,0008)	ImageType	DERIVED\PRIMARY\TOMOSYNTHESIS\NONE
(0008,0016)	SOPClassUID	1.2.840.10008.5.1.4.1.1.13.1.3
(0008,0018)	SOPInstanceUID	2.3.761.0.1.2402815.6.794.9587984116011978201
(0008,0020)	StudyDate	20150215
(0008,0023)	ContentDate	20150215
(0008,0030)	StudyTime	100538
(0008,0033)	ContentTime	100538
(0008,0050)	AccessionNumber	885B2947
(0008,0060)	Modality	MG
(0008,0070)	Manufacturer	HOLOGIC, Inc.
(0008,0090)	ReferringPhysicianName	BUTLER KEVIN
(0008,1030)	StudyDescription	MAMMO SCREEN BREAST TOMOSYNTHESIS BILATERAL

Identifiers

Dates

Name

PHI in DICOM Header: After MIDI

DICOM Tags

Search...

(Group,Ele...	TAG Description	Value
(0002,0000)	FileMetaInformationGroupLength	210
(0002,0001)	FileMetaInformationVersion	
(0002,0002)	MediaStorageSOPClassUID	1.2.840.10008.5.1.4.1.1.13.1.3
(0002,0003)	MediaStorageSOPInstanceUID	1.3.6.1.4.1.14519.5.2.1.8700.9920.443638053786977417456435
(0002,0010)	TransferSyntaxUID	1.2.840.10008.1.2.1
(0002,0012)	ImplementationClassUID	1.3.6.1.4.1.22213.1.143
(0002,0013)	ImplementationVersionName	0.5
(0002,0016)	SourceApplicationEntityTitle	POSDA
(0008,0005)	SpecificCharacterSet	ISO_IR 100
(0008,0008)	ImageType	DERIVED\PRIMARY\TOMOSYNTHESIS\NONE
(0008,0016)	SOPClassUID	1.2.840.10008.5.1.4.1.1.13.1.3
(0008,0018)	SOPInstanceUID	1.3.6.1.4.1.14519.5.2.1.8700.9920.443638053786977417456435
(0008,0020)	StudyDate	20350209
(0008,0023)	ContentDate	20350209
(0008,0030)	StudyTime	100538
(0008,0033)	ContentTime	100538
(0008,0050)	AccessionNumber	
(0008,0060)	Modality	MG
(0008,0070)	Manufacturer	HOLOGIC, Inc.
(0008,0090)	ReferringPhysicianName	
(0008,1030)	StudyDescription	MAMMO SCREEN BREAST TOMOSYNTHESIS BILATERAL

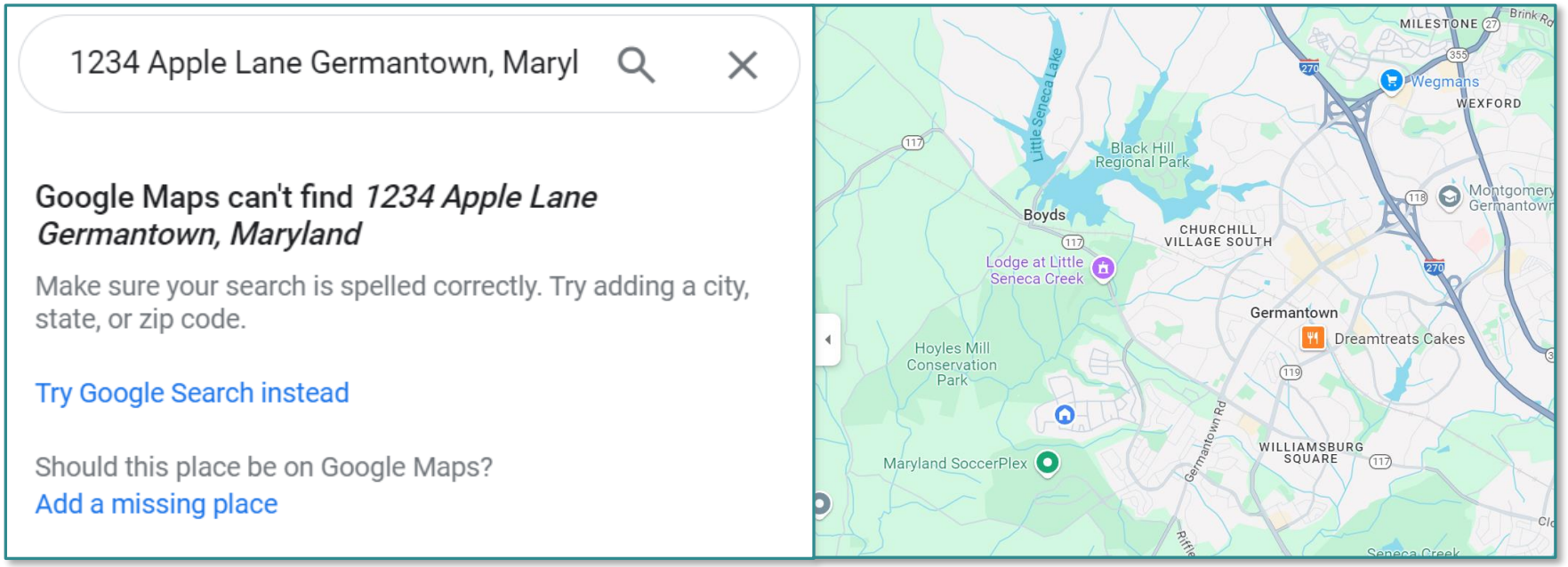
Identifiers changed

Dates randomly shifted

Name removed

MIDI Testing Misstep: Google Maps

- Synthetic PII including made-up addresses used to test MIDI
- Google algorithm could not find fake addresses in Google Maps
- The “missed” PII flagged as an error in MIDI pipeline

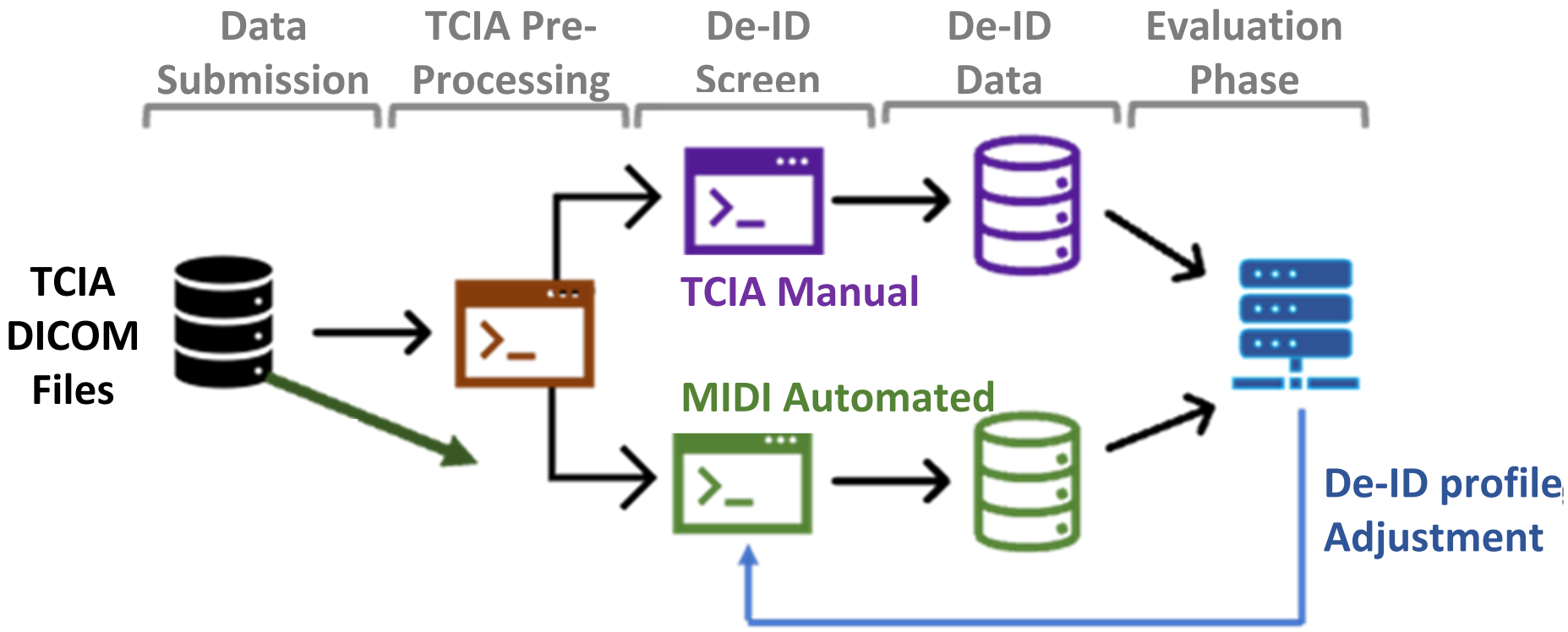


Address correctly
NOT FOUND
TO EXIST

NO MIDI
MASKING
of the synthetic PII

MIDI Pipeline: Phase 3

Real PHI to evaluate: The Cancer Imaging Archive (TCIA) manually curated DICOM images v/s MIDI automated



MIDI Benchmark Challenge

- **Goals**
  - Survey community de-identification tools
  - Develop a benchmark dataset and validation method to evaluate algorithms
- DICOM images with synthetic PHI divided into validation and test datasets
- 80 registered, 10 teams completed the test phase
- Sage Bionetworks was the challenge platform

MIDI Benchmark Challenge: Results

- 581,265 tracked actions at known locations
  - Changing identifiers, shifting dates, removing PHI
  - Actions also included not removing non-PHI
- Top 5 teams scored 99.87-99.93% correct
- No team led for all types of actions
- Possible to combine best practices from teams
- MIDI pipeline scored comparatively to the Top 5

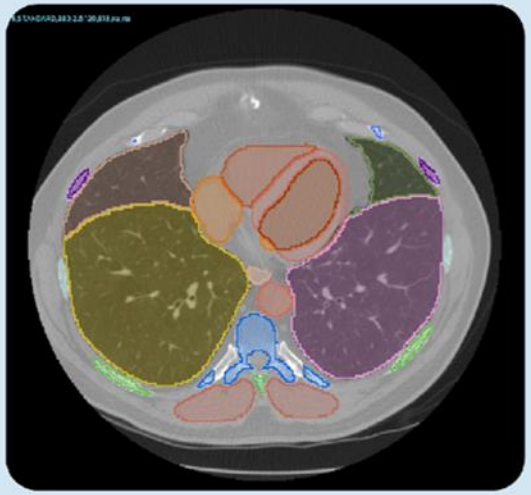
MIDI Components & Accomplishments

- **MIDI Datasets:** Medical images with synthetic patient identifiers [2020-2022], 2021 publication
- **Task Group:** Image de-identification guidelines and best practices [2022-2023], 2023 publication
- **Workshop:** Image de-identification in radiology and digital pathology [May 2023], reports 1 and 2
- **MIDI Benchmark Challenge:** Objective assessment of image de-identification tools [Oct 2024], results
- **MIDI Pipeline:** Scalable & AI-enabled image de-identification [2020 -], based on Google Sensitive Data Protection API

Model Hub (MHub)

- Platform for reproducible AI models
- Reproducible AI image processing
  - Harmonized input/output, containerized
- Deep Learning Models for feature annotation, prediction, and classification
- Models published in the literature
- Features generated by one model can be used as input for other models

IDC <-> MHub Virtuous Cycle

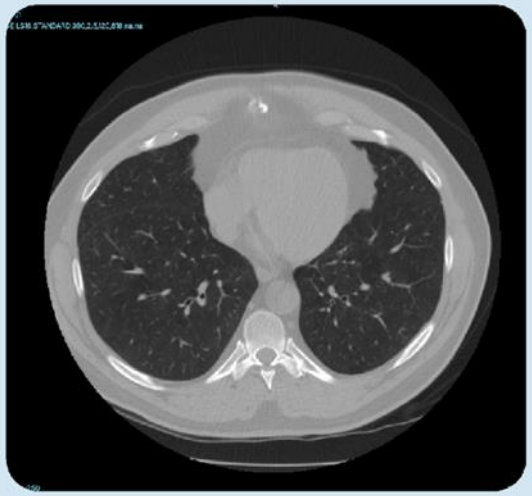


IDC

Image with No Features

MHub

Harmonized Model

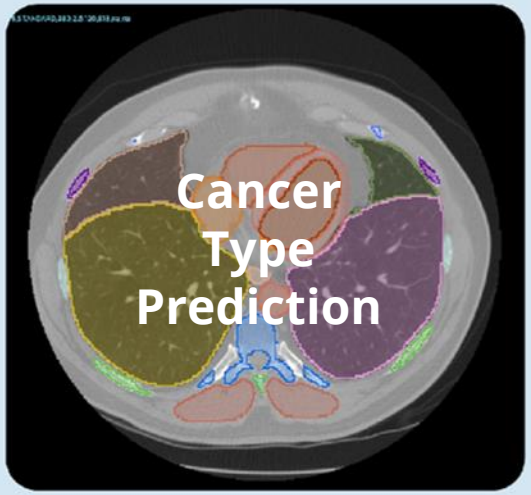


① An IDC image with no features gets fed into an MHub model

② The harmonized model generates new image features

③ The features are captured as part of the image in the IDC

IDC <-> MHub Virtuous Cycle

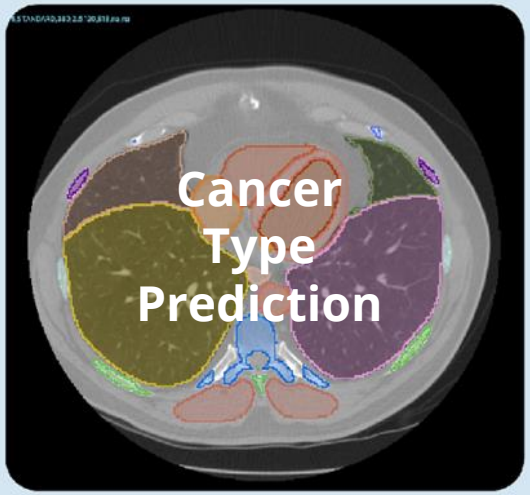


IDC

Image with Features

MHub

Predictive Model



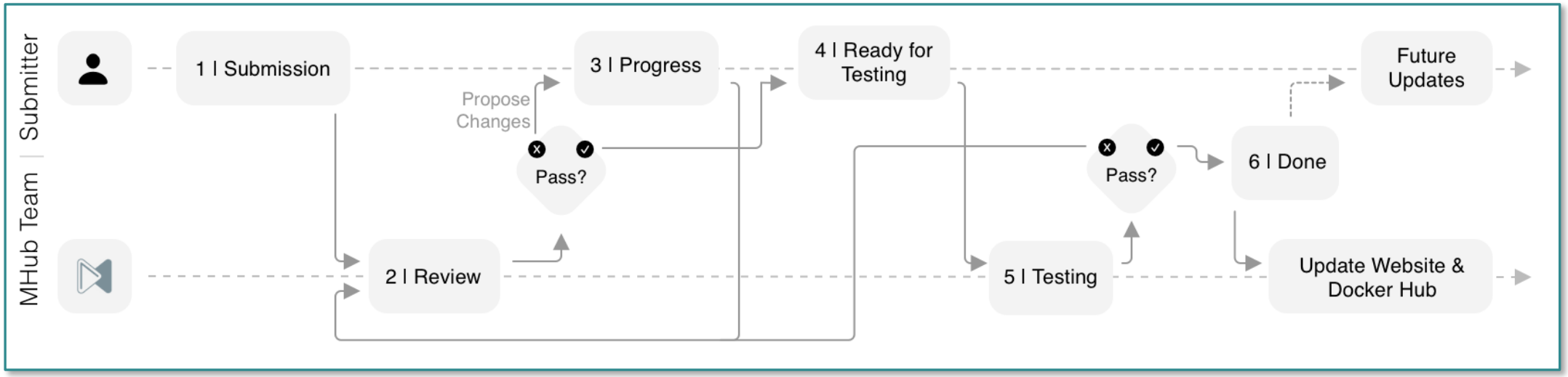
① An IDC image with features gets fed into a predictive MHub model

② The predictive model generates cancer type prediction

③ The cancer type prediction is stored as metadata in IDC

Self-Serve MHub & IDC AI annotations

- Planned Self-service capability for MHub
  - Step by step guides, tutorials, Github framework
  - Based on published AI models
- Super-charging the IDC <-> MHub Virtuous Cycle



Medical Image Synergistic Workflow

- Medical Image De-Identification (MIDI)
  - Users can de-identify DICOM images
  - CRDC to validate de-identification at submission
- Imaging Data Commons (IDC)
  - Stores de-identified DICOM images
  - Download and/or analysis with tools in the cloud
- Model Hub (MHub)
  - Harmonized AI model repository, accepting DICOM as input and producing DICOM and other outputs
  - Virtuous cycle: MHub AI models run on IDC datasets produce annotation datasets, stored back in the IDC for improved data

Acknowledgements

NIH Staff

- **FNL:** U.Wagner, L.Pei, T.Pihl
- **NIH/NCI:** K.Farahani, L.Finkelstein, K.Seshadri, S.Pan, T.Davidsen, E.Kim, C.Hammond, N.Weber

MIDI Team

- **UAMS:** F.Prior, M.Rutherford, K.Smith, J.Underwood
- **Pixelmed:** D.Clunie
- **Ellumen, Inc.:** Q.Pan, S.Gustafson
- **Deloitte:** B.Kopchick, J.Klenk, L.Opsahl-Ong, K.Johnson, T.Do, S.Boppana
- **Google:** B.Lou, C.Corman, D.Belardo, D.Hawkins

IDC/MHub Teams

- **Brigham and Women's/Mass General:** A.Fedorov, R.Kikinis, C.Ciausu, D.Krishnaswamy, V.Thiriveedhi, H.Aerts, C.Bridge, L.Nürnberger, D.Bontempi, B.Suraj Pai
- **ISB:** B.Longabaugh, B.Clifford, S.Paquette, G.White, D.Gibbs, I.Shmulevich
- **GDIT:** D.Pot, F.Seidl, P.Gundluru
- **Isomics Inc:** S.Pieper
- **PixelMed:** D.Clunie
- **Radical Imaging:** R.Lewis, I.Octaviano
- **Fraunhofer MEVIS:** A.Homeyer, D.Schacherer