6579

Accelerating de-identification of images with cloud services to support data sharing in cancer research

Ben Kopchick*[1], Laura Opsahl-Ong[1], Qinyan Pan[2], Michael Rutherford[3], Ulrike Wagner[4], Bhavani Singh[1], Scott Gustafson[2], Fred Prior[3], David Clunie[5], Juergen Klenk[1], Keyvan Farahani[6]

Introduction

Imaging databases such as NCI's Imaging Data Commons (IDC) stand to benefit from automated de-identification. Protected Health Information (PHI) and Personally Identifiable Information (PII) can be embedded in these images themselves (pixel data) as well as in the metadata (DICOM header). Repositories, like The Cancer Imaging Archive (TCIA), utilize the DICOM standard (specifically PS3.15 Appendix E) as a baseline for which elements are labeled as PHI/PII. This standard provides guidelines for selecting elements which are expected to contain PHI/PII, so that they can formulaically be removed (or anonymized). For medical data a balance must be struck between removing information which contains sensitive information (to ensure privacy) versus retaining information that is critical within a certain research context.

With an ever-growing volume of imaging data, a manual approach to de-identification becomes infeasible, expensive, and prone to error. An automated system that employs ML/AI can help improve the de-identification accuracy and expedite the process allowing image data to be shared amongst researchers sooner.

One solution for de-identification of medical images is the Google Cloud Platform's (GCP) Healthcare API. Based on Google's Data Loss Prevention service, it offers a configurable system that is scalable for large and growing datasets. Additionally, a 'human-in-the-loop' can be included for spot-checking the de-identified datasets. This combination of automation and human detection would improve accuracy and speed of the de-identification process.

Methods and Materials

We established the infrastructure for this project on NCI's Cloud 2 (GCP) environment, which includes the infrastructure seen in Figure 1. It involves (1) loading images into a Cloud Storage Bucket, (2) transferring the data to a DICOM store using the GCP Healthcare API, (3) running the de-identification service with appropriate configuration flags, (4) moving images back to a storage bucket, and (5) analyzing the results. De-identification is performed using an alpha release of GCP's Healthcare API.

A dataset containing 216 patients and 23,921 images was prepared to test the de-identification algorithm by placing synthetic PHI in both DICOM headers and pixel data. The synthetic data matched real data seen during the curation of data at TCIA. Accuracy of the MIDI pipeline was measured against TCIA's standard tools and procedures for de-identification. Measures included correct detection of all PHI data and correct action taken (e.g., remove, encrypt, or otherwise obscure). We also measured throughput of the pipeline.

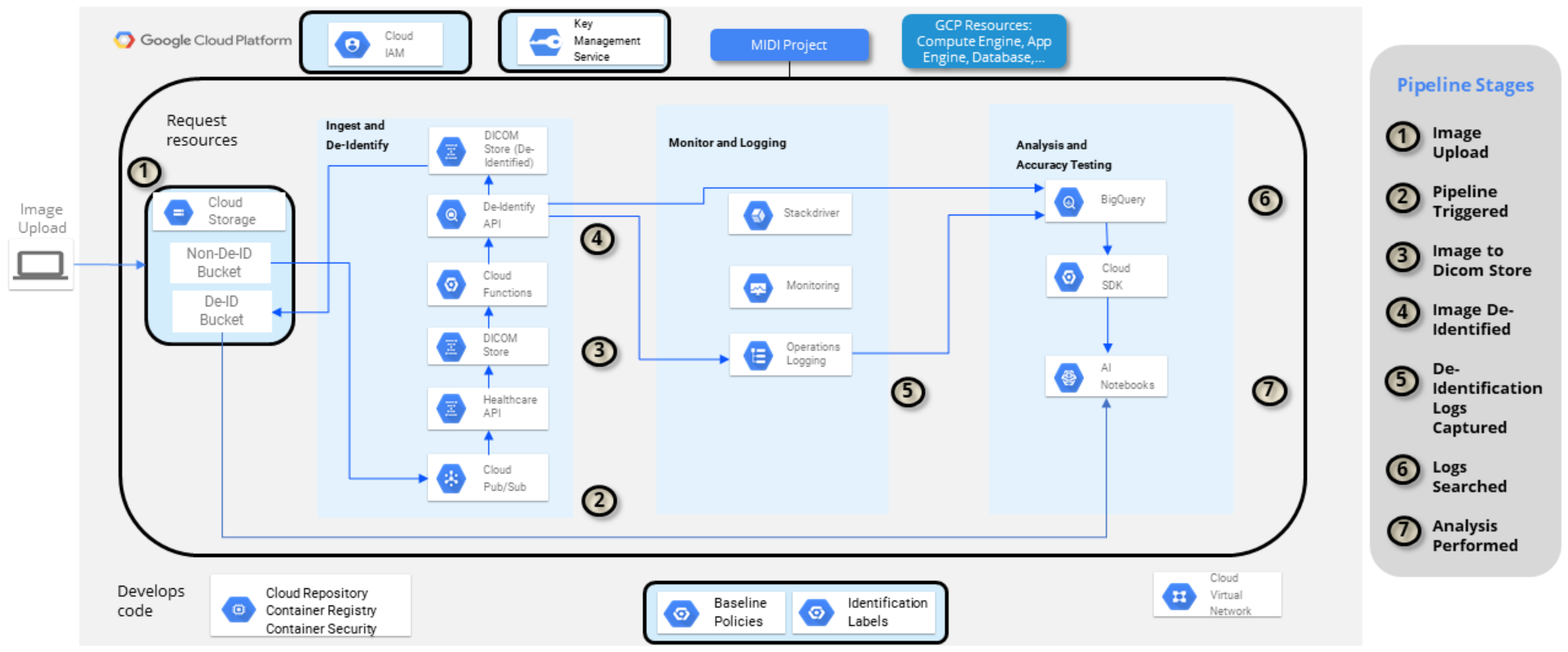

Figure 1. De-identification pipeline architecture in GCP.

Results

The MIDI pipeline's accuracy for DICOM headers was 98.7% (Table 2), accurately detecting dates, addresses, phone numbers, unique identifiers, names, and other common PHI. An example of input and output values can be found in Table 1. The most common PHI failed to remove included names, dates in string data types, patient IDs, and abbreviated institution names. Private Creator data elements were consistently failed to be retained. UIDs were correctly replaced. PHI burnt-in the pixel data was successfully detected and removed, with one false positive. An example of de-identified pixel data is shown in Figure 2.

Throughput was measured at 22.0 images per second over 10 runs. This means for the 23,921 images it took on average 18 min 7 sec. Throughput is dependent on multiple factors including latency and time of day due to availability of resources.

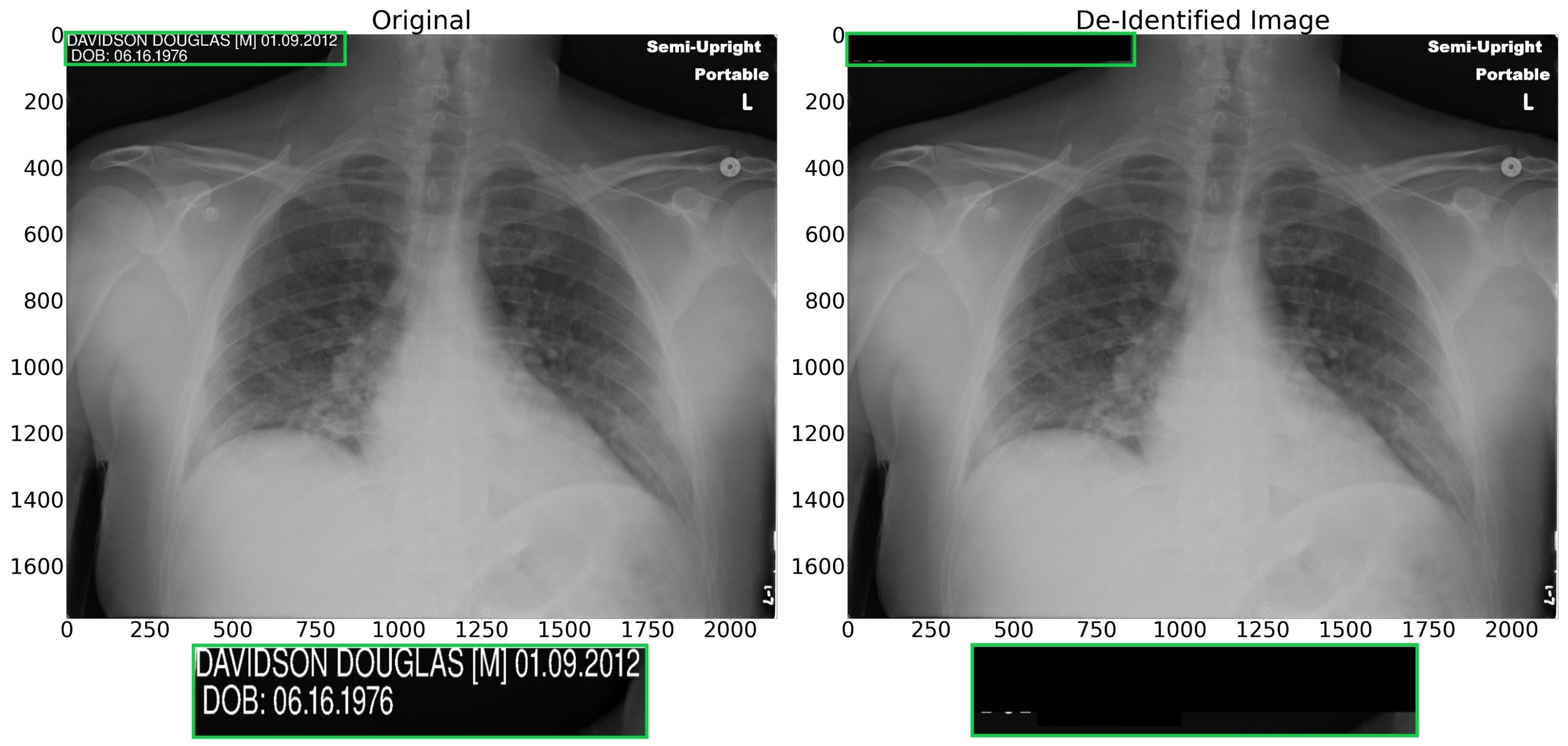

Figure 2.(Left) Example of an image containing sensitive and non-sensitive text. (Right) Image post de-identification with sensitive text removed and non-sensitive text remaining.

Action	Tag Name	Input	Output
Text Removed	Study Description	MRI Prostate W WO Contrast for Madison	MRI Prostate W WO Contrast for [Person_Name]
Date Shifted	Study Date	19930722	19930218
Text Retained	Additional Patient History	SIEMENS MEDCOM HEADER	SIEMENS MEDCOM HEADER

Table 1. Example of DICOM header data and it's de-identified output

Discussion

While the de-identification pipeline performed well around actions where text needed to be retained, dates shifted, and text removed on pixel data, there are still issues regarding necessary text being removed. Most of these issues were due to Patient Ids not being properly recognized and removed. While 'Patient ID' data elements were removed 100% of the time, patient IDs that appeared in free text data elements such as 'Study Description' were difficult for the algorithm. Another current area that needs improvement is the recognition of non-western and atypical names as well as names in atypical formats. This formatting includes names that contain an underscore, such as 'A_John Doe', where the 'A_John' is not correctly identified as a name. In a production version, tags and strings that have been shown to be problematic would be identified for a 'human-in-the-loop' for manual correction.

Action Taken	Percent Correct
Text Retained	99.2%
Text Not Null	100%
Pixels Hidden	100%
Date Shifted	98.3%
Text Removed	84.7%
Total	98.7%

Table 2. Percent correct of de-identification by action that was supposed to be taken per unique tag.

Conclusions

We demonstrate the current capability and performance of automated cancer image de-identification. Our results show that while full automation is within grasp, a semi-automated pipeline is now feasible. A human expert in the loop can be used for final verification. This will lead to a much-needed acceleration of cancer image de-identification, to handle the rapidly growing volume of cancer image data and provide rapid data access to accelerate research. Future work will focus on including pre- and post-processing tools to aid the human expert in the loop, such as identifying and flagging questionable images for manual review.

[1]Deloitte Consulting LLP
[2]Ellumen, Inc.
[3]University of Arkansas for Medical Sciences
[4]Frederick National Laboratory for Cancer Research
[5]PixelMed Publishing
[6]National Cancer Institute

*Contact Information:
bkopchick@deloitte.com