LifeImage Google Cloud

WHITEPAPER

The essential features of a clinical data and diagnostic imaging de-identification process that can operate at scale

Contents

Introduction

There is growing industry pressure to more readily share clinical trial data in order to accelerate medical insights. However, a vast amount of medical image data collected in the course of clinical trials has been unusable for research and innovation. One common reason is the lack of a reliable method of removing personally identifiable information (PII) and personal health information (PHI) from medical image datasets, a process known as de-identification.

De-identification is difficult to perform at enterprise scale for all forms of clinical data, but is particularly difficult for medical images, due to the complexity and lack of standardization in DICOM metadata.

Google Cloud created a de-identification application programming interface (API) as one of the tools available in its Cloud Healthcare API, which are standards-based APIs created to power actionable healthcare insights for security and compliance-focused environments.

Life Image evaluated the de-identification capabilities of the Google Cloud Healthcare API against a number of open source, proprietary and commercial de-identification programs and found that the Google Cloud Healthcare API outperformed other tools in the market in its accuracy at de-identifying PHI contained in imaging pixels, metadata tags and associated notes.

Life Image further augmented the performance of the Google Cloud Healthcare API with a value-add service that uses a combination of machine learning and human validators to achieve accuracy levels consistent with some of the most stringent compliance requirements and beyond those operating in healthcare today. Combined, the Life Image and Google Cloud Healthcare API can help protect patient information and promote accelerated research development.

The Market Need for De-identification

Collecting medical images for clinical trials is an expensive and time consuming process. These medical images are typically used to answer the specific clinical question in the trial and not used further. By enabling the re-use of this data after the trial has ended, future medical insights can be gained faster and with reduced cost. Removing all personal identifying information from these medical images, a process called de-identification, helps to enable the use of this data beyond the scope of the individual clinical trials.

Integrating these diagnostic images into larger datasets would permit the use of big data techniques, such as machine learning, to make significant medical advances. Additionally, patients who participate in clinical trials typically want the data that they generate to continue helping others after the trial has ended.

The problem of inaccessible medical data is widely known. There is growing pressure from regulators, pharmaceutical companies, medical device manufacturers, AI companies, and academic medical centers to make medical imaging data more readily available for analysis. For example, as of January 1, 2019, the member journals of the International Committee of Medical Journal Editors (ICMJE), which includes such journals as the New England Journal of Medicine and Annals of Internal Medicine, require authors to disclose their data sharing plans before commencing a clinical trial.[1]

Organizations with existing relationships and data use contracts, such as a HIPAA Business Associate Agreement (BAA), can share data without fully removing all personally identifiable information (PII) and protected health information (PHI). However, if these data are to be shared more broadly and combined into larger datasets, there must be a way to flexibly remove or redact such identifying data. This removal of identifying or protected information is called de-identification.

Specific Problems Faced in De-identifying Diagnostic Imaging Data

Diagnostic images are a valuable form of clinical data, containing both structural and functional information about disease states. Progress of disease can be measured, tracked, and communicated with a high level of detail.

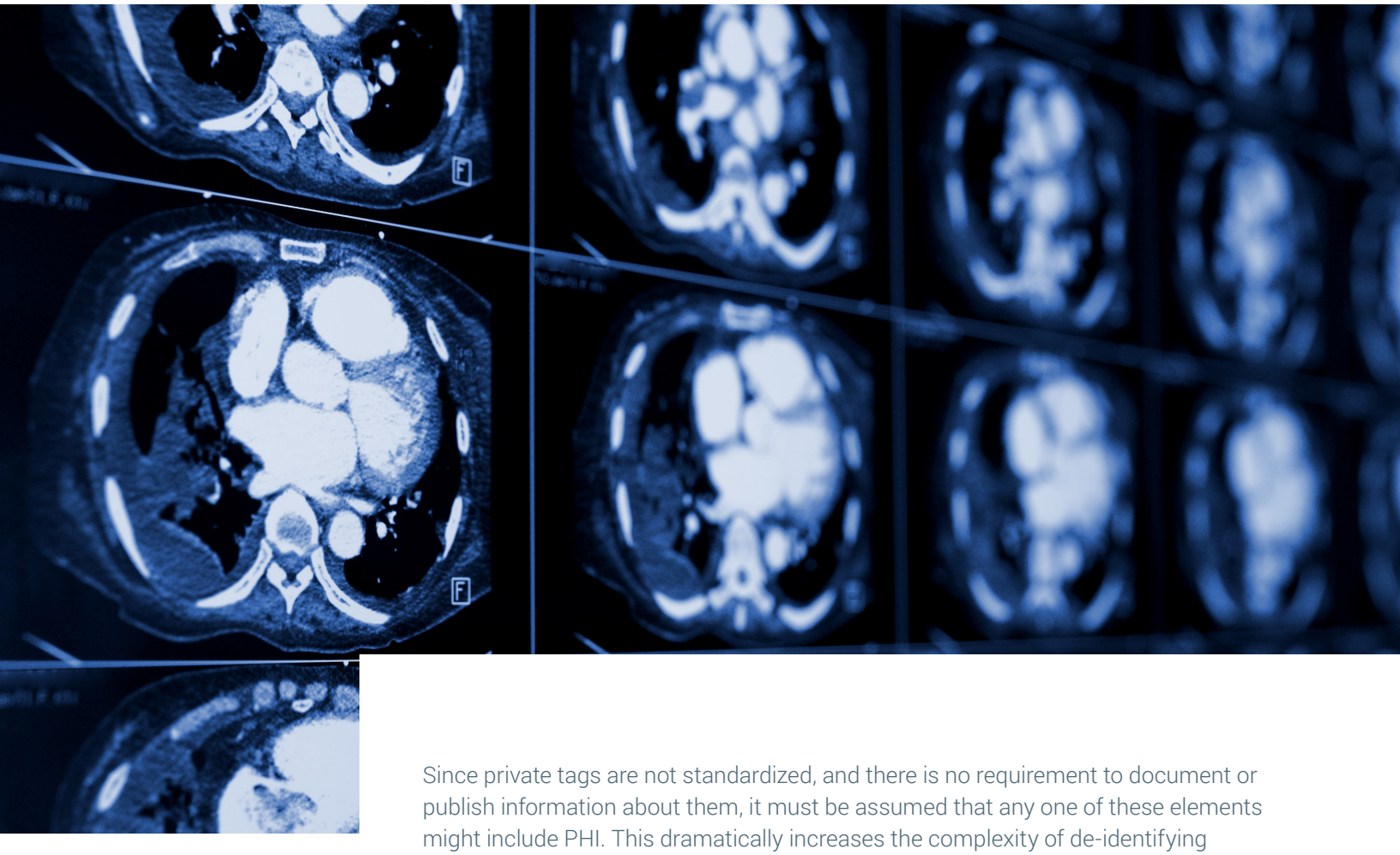
De-identifying diagnostic images without removing relevant clinical information can be a significant challenge. Until now, it has been a difficult, slow, and expensive process that required specific knowledge of the manufacturer, device, and modality used for each set of images.

Digital Imaging and Communications in Medicine (DICOM) is the international standard for managing, communicating, and storing diagnostic image data. It was developed to enable interoperability between diagnostic imaging manufacturers, healthcare facilities, and clinical laboratories. In addition to the image itself, a DICOM file contains metadata elements ("tags") that encode large amounts of information about the acquisition of the image, the patient, and other context.

While the DICOM standard has hundreds of "public" tags that define standard image acquisition parameters, it also allows each manufacturer to use "private" or "custom" tags for encoding proprietary or as-yet-undefined acquisition parameters. For example:

- Instrument manufacturers may use private tags to include their proprietary hardware and software settings that define image acquisition.
- Each imaging modality -- such as chest X-ray, CT scan, MRI, etc. -- can also include a different set of private tags to store modality-specific information.
- Each department, site, and/or institution may use different settings and customizations for their instruments depending on their specific needs. They might use private tags to store information about these settings.

[1] http://www.icmje.org/news-and-editorials/data_sharing_june_2017.pdf



Since private tags are not standardized, and there is no requirement to document or publish information about them, it must be assumed that any one of these elements might include PHI. This dramatically increases the complexity of de-identifying DICOM resources.

There are a number of additional issues that can cause further complications. Clinical sites may attach screen captures that include documents with PHI in them. Some modalities can put patient demographic information into the actual pixel data. Any element that incorporates user input as a text string could include PHI, such as the name of the referring physician, whether or not that element is intended for that use. And manual annotations can appear on the image itself, outside of the DICOM encoding entirely.

As a result, PII and PHI tend to be included idiosyncratically in the images, in a wide range of locations and formats, often not explicitly labeled or documented.

Shortcomings of Current Solutions

The most reliable methods of de-identification up until now are specific to certain purposes or uses and require significant amounts of skilled human involvement. These efforts take resources away from the actual clinical purpose of the data being collected.

These methods can de-identify specific types of datasets, for specific site locations, generated by specific devices, but they are not reproducible at enterprise scale.

De-identification methods used in the past were also typically restricted in the volume of data they could process without unacceptable levels of error. Overly aggressive de-identification methods could also cause significant amounts of clinically relevant data, such as patient demographics (age, gender, etc.), to be stripped out. Simply removing all possibly identifying information damages the integrity of the clinical data contained in the diagnostic images.

This reveals another problem faced by these customized, relatively slow solutions: they are hard to validate. Large scale validation with human intervention is slow, tedious, and restricted to small data sets. As a result, the de-identified datasets that are currently shareable and available tend to be both small and old, at a time when large, current datasets are what is needed.

As a result of these challenges, even de-identification procedures that have proved themselves in a specific context have a low level of trust among other potential users. They are aware that the possible consequences of an inadvertent release of PHI might include significant regulatory, civil, and even criminal penalties. These penalties can be imposed not only by the Department of Health and Human Services (HHS), but also by various state attorneys general. These risks are borne by the original researchers who collected and submitted the data as well as by their institution and the institutions that collected, stored, and disseminated the data. Thus, the upside for anyone with clinical trial data to share is limited, while the potential downside seems significant. The effort and cost of sharing the data combined with the risks, real or perceived, of any incomplete de-identification discourages data sharing.

A low-overhead, fast, validated way of anonymizing clinical data would significantly shift the incentives in favor of sharing, which would have a large effect on the availability of clinical data for all potential users.

What an Ideal De-identification Solution Would Look Like

There are two fundamental steps involved in anonymizing healthcare data, and both require a significant amount of computation. They are detection, finding and identifying the PHI within a resource, and transformation, where that PHI is either deleted, replaced, or encrypted, depending on the intended use of the data.

But those two steps are not useful without verification, an assurance that PHI is actually thoroughly removed. Just developing a reliable model requires large amounts of data, on the order of millions of clinical images. Verification takes even more data, as well as an expertise in the idiosyncrasies of clinical imaging, to tune the model and prevent overfitting.

Life Image evaluated a number of tools including open source, proprietary and commercial de-identification programs against the ability to detect and transform. The company selected the Google Cloud Healthcare API as the preferred tool after extensive analysis.

Life Image has used its own datasets to independently validate the performance of the Google Cloud Healthcare API for clinical image data. Life Image confirmed that the redaction of text and images in the de-identification process performed as well as, and in many cases outperformed, existing tools with faster speeds and higher accuracy than were previously possible.

Google Cloud used a wide range of clinical and non-clinical imaging to develop the Cloud Healthcare API's de-identification capabilities, focusing on increasing efficiency and scale without compromising accuracy. As a result, based on Life Image's assessment, the accuracy of the Healthcare API's de-identification meets and often exceeds that of any other existing automated tool, while significantly increasing the scale of datasets it can de-identify.

Life Image has evaluated the Google Cloud Healthcare API with its own diagnostic imaging data. In addition, Life Image developed a value-add service layer that uses a combination of machine learning and human validators to further augment the Cloud Healthcare API's performance in order to achieve accuracy levels consistent with some of the most stringent compliance requirements and beyond those operating in healthcare today.

Detection

When performing de-identification, the Cloud Healthcare API first inspects the medical images to identify the locations of potentially sensitive data, such as patient names, birthdates, physician names, and so forth. These can be in various places, in a wide variety of formats, as described above.

After data has been identified as potentially containing sensitive information, it is then subjected to a combination of rules-based, heuristic, and machine learning (ML) analysis to determine if there is PII / PHI that needs to be removed. Rules-based methods are fast and derived from knowledge of the structure of specific data types, such as email addresses and phone numbers. Heuristic methods can further improve accuracy without sacrificing speed. ML methods are the most complex and computationally intensive, but they can be the most effective method for detecting PII and PHI when confronted with undefined or poorly defined data formats. The Google Cloud Healthcare API flexibly uses a combination of these methods, allowing it to efficiently and effectively process large datasets.

Detection has long been the difficult analytical problem in unstructured healthcare data, particularly clinical image data. Existing de-identification tools are primarily rules based and a few are heuristics based. This limits their ability to de-identify

medical data with known and defined formats. The innovations in the Google Cloud Healthcare API are more accurate heuristics and the addition of a machine-learning algorithm that works effectively to de-identify medical images in a variety of formats. Few clinical datasets are clean and defined, and this is where the Google Cloud Healthcare API truly demonstrates value.

Transformation

After the PII / PHI has been detected, the next step is transformation. Transformation can be a simple deletion of the identifying data, but the Google Cloud Healthcare API also supports more complex transformation methods such as date-shifting, substitution and encryption.

Date shifting involves changing all dates in a medical imaging file by the same randomly chosen amount, usually between one and 100 days. Dates in a medical file can have clinical value, but can also include identifying information. Shifting all dates in a single file by the same amount allows for the removal of identifying information, such as birthdays, while preserving the clinical value of the chronological order and relative time between events.

Another transformation option is the substitution of “dummy data” in the place of identifying information, such as replacing the actual patient name (e.g. “John”) with a realistic-looking substitute (e.g. “Daniel”). This helps maintain human-readability of data while protecting the privacy of the original patient.

Encryption allows for replacing potentially identifying information with encrypted codes, which can be decoded with a “key”. This “key” can be retained by the original dataset owner or even a third party, who can be contacted if there is a medically necessary reason to decrypt the data.

Transformation is essential for data integration and data management because different use cases require different levels of information. The Google Cloud Healthcare API’s de-identification tool is capable of a wide range of data transformations, flexibly producing a level of de-identification appropriate to the need and intended use.

Researchers will no longer have to spend scarce resources developing their own de-identification methods. Instead, they can devote themselves to their research goals without being limited to older, smaller datasets that limit the ability to reach reliable conclusions.

The Human Element

Even with the power and scope of the Google Cloud Healthcare API de-identification, it is just one step in preprocessing clinical image datasets. Data normalization, cleaning, and scaling must still be performed. Specific use cases may require additional configuration, and the results may need to be further examined to validate the performance and accuracy of the de-identification.

The benefits of the Google Cloud Healthcare API are automation, scale, quality, and cost effectiveness, all benefits previously unavailable for de-identifying healthcare data. After this de-identification procedure is performed, it is a good practice to have a human expert to validate the results. This expert can examine a subset of the output to verify that the data has been de-identified with an acceptable level of quality and accuracy, and that the procedure for that dataset produced the expected results. Depending on the specific use case, one may even choose to validate anything from a representative subset of the output, to the entire output dataset.

Availability of Large De-identified Clinical Image Datasets from Life Image

With its 11-year history as a leader in medical image exchange, Life Image has extensive experience with every aspect of imaging data, particularly the complex and specific ways diagnostic images can conceal PII and PHI. It has used this knowledge to develop a service layer of specific machine learning algorithms that further improves the performance of the Google Cloud Healthcare API in de-identifying clinical images.

Skilled clinical validators work in conjunction with this service layer to de-identify large clinical image sets at an unmatched level of accuracy. Once Life Image has used the Cloud Healthcare API with Life Image's additional service layer and human validators on its large clinical image datasets, Life Image can offer datasets that are virtually free of errors. The datasets will be recent, large, and varied, supporting a wide range of research.

All of the complex processing, cross-checking, and testing is performed by Life Image, and the results are large clinical image datasets that are completely de-identified and are usable for clinical research with the assurance that no PII or PHI is contained in them. Researchers will no longer have to spend scarce resources developing their own de-identification methods. Instead, they can devote themselves to their research goals without being limited to older, smaller datasets that limit the ability to reach reliable conclusions.

Conclusion

Google Cloud emphasized scalability, cost effectiveness, and performance when designing the de-identification capabilities of the Google Cloud Healthcare API.

The de-identification available with the Cloud Healthcare API will make it significantly easier to share clinical data.

Life Image used its own large clinical image datasets to validate the Google Cloud Healthcare API. Life Image's additional protocols have refined their use of the Cloud Healthcare API with DICOM images. Used in conjunction with the Cloud Healthcare API, these protocols further increase the accuracy of the de-identification procedure.

Life Image is currently working to leverage this new approach to create large de-identified clinical image datasets useful for training machine learning models, understanding drug effectiveness on patient subgroups, comparing imaging devices from various vendors, and a range of other clinical studies.

LifeImage® Google Cloud

About Life Image

Life Image is the world's first truly interoperable clinical information exchange network. Our platform ecosystem that benefits every player, including patients, across the healthcare ecosystem, enabling shareability and timely access to patient information at the point of care.

One Gateway Center
300 Washington St, Suite 200
Newton, MA 02458

Tel: 617.244.8411 | Fax: 617.244.8611

lifeimage.com

Google Cloud Healthcare API

Cloud Healthcare API bridges the gap between care systems and applications built on Google Cloud. By supporting standards-based data formats and protocols of existing healthcare technologies, Cloud Healthcare API connects your data to advanced Google Cloud capabilities, including streaming data processing with Cloud Dataflow, scalable analytics with BigQuery, and machine learning with Cloud Machine Learning Engine. In addition, Cloud Healthcare API simplifies application development and device integration to accelerate digital transformation and enable real-time integration with care networks.

©2019 Life Image, Inc. All rights reserved. Life Image and the Life Image logo are trademarks of Life Image, Inc. All other brands may be the trademarks of their respective owners.