Canopy

CASE STUDY

Canopy's Auto Review Completes "Impossible" PHI Data Mining Project

OVERVIEW

A large hospital network experienced one of the biggest protected health information (PHI) breaches of the year, with over 6,600 compromised PDFs.

The documents were lengthy and densely packed with PHI, causing a data recognition and extraction problem. And because these were daily billing reports, patients' information was frequently duplicated; the same person often could appear thousands of times.

The hospital network's review management company needed to quickly discover who was impacted and what PHI was compromised so that they could respond in compliance with the HIPAA Breach Notification Rule.

SOLUTION

Canopy's advanced detection algorithms first identified each PII/PHI element. Auto Review linked patients' names to their procedure dates, medical codes, insurance information, and other personal data. The only human effort required for review was for QA.

Our machine learning models consolidated the entities into a list of unique patients, maintaining links to each person's data elements.

RESULTS

Canopy completed this otherwise impossible project in just 15 days. We transformed the data into a structured format, scanned it, and generated a final consolidated entity list containing just 3 million unique patients.

Canopy

The Clock Starts the Moment an Incident Occurs

When sensitive data is compromised, HIPAA-covered businesses in the United States must quickly assess whether personally identifiable information (PII) or protected health information (PHI) were disclosed. If so, they are legally obligated to notify affected individuals within 60 days following the discovery of the breach.

Problematically, many incident response (IR) teams still do this using tools and workflows that aren't specifically designed for IR data mining. This results in slow, manual PII/PHI review and notification list generation that cannot handle large data sets in the required timeframe — putting the business at risk of hefty fines and reputational damage.

This is the challenge that a review management company faced when one of their customers, a large hospital network, experienced a significant PHI breach. The PII and PHI (including account and claim numbers, procedure and diagnostic codes, and insurance plan information) were stored in tables within more than 6,600 PDFs generated from standardized Crystal Reports, some of which contained over 180,000 rows of information. The compromised PDFs were daily billing reports, so patients were frequently duplicated — the same person could appear thousands of times with varying PII and PHI.

Using their typical ediscovery-based software and workflows, the IR team would have needed to manually copy each entity's name and information from the PDF tables into a spreadsheet. Given the sheer number of documents and the frequent duplication of patients in this project, this simply was not possible. The 6,600+ PDFs ended up containing 4.28 billion entities. Assuming the average reviewer can transcribe 120 entities per hour, it would take approximately 35,672,000 man hours to copy the data over.

Data Mining Done Right with Canopy

Recognizing this to be an impossible project using their typical methods, the IR team turned to Canopy's Auto Review, which uses AI to automate the process of connecting detected PII/PHI to people in standardized data sets.

The application processed the data, scanning, mapping, and classifying the PII/PHI. Then, Auto Review linked patients' names to their procedure dates, medical codes, insurance information, and other personal data. The only human effort required for review was for QA.

Finally, to enable the hospital network to fulfill their legal obligation to notify in a timely manner, the incident response

By the Numbers

6,600
Crystal Reports

4.28 billion
entities

60 days
to respond

Ediscovery Methods

35,672,000 man hours
simply to convert the data to a structured, scannable format

Canopy's Solution

15 days
to complete the entire project

team needed a consolidated list of who was affected by the breach and what data was compromised. With just a few clicks, Canopy solved for the frequent duplication of patients using its AI-powered entity deduplication. This enabled the IR team to quickly consolidate their list by over 99% — from 4.28 billion down to 3 million unique people.

After the automated review and deduplication were complete, Canopy exported a consolidated entity list that was formatted to the hospital network's specifications.

Canopy Made the Impossible Possible

Using old-school ediscovery-based methods for incident response data mining results in significantly longer timelines and costlier service than Canopy's purpose-built workflows and AI-powered software.

In this case, the "old way" of approaching data mining wasn't even a realistic option — it would have taken a team of 100 hourly-paid reviewers over 120 years just to transform the data into a structured format, assuming they worked nonstop eight hours per day, seven days per week.

> "It was not humanly possible for our team to do this — it would have taken a couple hundred reviewers years to complete this project. We can't even fathom the cost savings. Canopy [Auto Review] made the impossible possible."
>
> **— Project Lead**

Canopy completed the entire project from start to finish in just 15 days using:

- Auto Review's AI linking PII/PHI to people.
- A streamlined deduplication process.
- A final entity export of 3 million unique patients — 99.9% smaller than the original entity list.

Canopy's Auto Review made this project possible, ultimately enabling the hospital network to alert impacted individuals quickly and in compliance with HIPAA data breach requirements.