Testing with De-identified Protected Health Information (PHI)

Wade Donahue, Lead EA Architect 7/27/16

It seems all too frequent, another data breach involving Protected Health Information (PHI). This has not only resulted in a greater focus on data protection by regulatory agencies, but consumers are now demanding greater levels of assurance that their personal information is being protected. As a result, many organization are making significant people, process, and technology investments to reduce the risk of a data breach. One way to mitigate risk is to test with production data that has been de-identified.

De-identification is a process of taking data fields that contain PHI and either removing them or replacing the contents with other realistic data. For example, phone numbers are replaced with fictitious phone numbers. The HIPAA Safe Harbor method has become a benchmark for the field types to be de-identified. These field types include name, address, phone numbers, Social Security Numbers, etc.

Moving to testing with de-identified data starts with communicating that goal to all levels of the organization. Make sure to communicate early and often using a variety of media. Now that everyone understands the goal, you need a software solution that provides a variety of methods for de-identifying data. Unless you have very clean data, make sure any solution chosen has custom logic capabilities to enable conditional de-identification.

As you are looking at solutions for de-identification, you'll find many of them provide capabilities to perform scans of your data and identify candidate data fields for de-identification. For example, the scan will identify a continuous string of nine numbers as an SSN. This type of scan only gets you started, as the software doesn't know what custom fields you may have created in that could identify an individual. This is where data owners and subject matter experts come in to play as they need to verify and augment the scan findings. By the way, did I mention communicating early and often? Let them know this is coming. The results of the verification process provides you with the information you need to configure the de-identification software.

Now that we have the software and have it configured, we are ready to start de-identifying data. It is important to allocate sufficient time to test the data de-identification process and applications because you will have unanticipated data issues to deal with. Unless you have very mature and automated test practices, be sure to allocate time in you plan to address you test case inventory to align the scenarios with the newly created set of de-identified data.

Moving away from testing with production data to testing with de-identified data is not only a technical problem, it requires a major culture shift for everyone involved in the development, testing and support of your applications. By using de-identified data for testing you have taken one step in reducing the risk of a data breach and meeting the expectations of regulatory agencies and healthcare consumers.