# 2022 NPCR NEBRASKA SUCCESS STORY

Nebraska Cancer Registry; Qianru Wu, Veronica Boudreaux, Christina Hiller, Mark Watson, Lifeng Li, Marissa Ayotte, Christophe Irumva, Mary Mesnard

## *De-identification Tool for HL7 Formatted Electronic Pathology Report*

## National Program of Cancer Registries
## SUCCESS STORY

### SUMMARY

The Nebraska Cancer Registry (NCR) and Tanaq Support Services, LLC (Tanaq), the contracted company with the Centers for Disease Control and Prevention (CDC) to implement the Childhood Cancer Survivorship, Treatment, Access, and Research (STAR) project, collaborated in developing a de-identification tool for Health Level Seven (HL7) formatted electronic pathology (ePath) report to assist in the development of the National Program of Cancer Registries - National Oncology rapid Ascertainment Hub (NPCR-NOAH). The tool automatically masked 18 types of identifiers of the individuals, relatives, employers, or households, defined by the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. It greatly increased the efficiency of de-identification compared to manual work.

### CHALLENGE

Developing and testing new software requires a vast amount of sample data. Sample data resembling real data as closely as possible with the identifying information masked per HIPAA rules. However, it usually takes 10-30 minutes for a professional to manually de-identify one HL7 message, depending on the length and complexity of the message. In Nebraska, the monthly volume of the incoming ePath reports is between 200 and 450 (on average 6-15 reports per day). Therefore, manual de-identification can take hours of work every day and will be a huge burden even for a relatively less populated state like Nebraska.

### SOLUTION

The NCR and Tanaq collaborated in developing a de-identification tool to generate sample HL7 messages from real HL7 files. First, we identified 49 data fields in an HL7 message that needed to be de-identified, per the "Safe Harbor" method of the HIPAA Privacy Rule. Then we wrote a Java program that read the HL7 file by its structure, looked for identifiable data fields, and replaced identifiers with pseudo data. Among all 49 fields, the most challenging one was OBX-5, which recorded the narrative of the pathology report and thus was important for abstraction. Unlike the other 48 fields, which had very strict rules governing data content and data length, OBX-5 was free text, meaning that any text could be put in this field in any layout. Identifiers, such as the patient's name or provider's name, were always embedded within diagnostic content in OBX-5. The ideal solution was to retain only the diagnostic text while masking all identifiers. However, the implementation was impossible even in today's technology stacks because the content of identifier was unpredictable – it could be anything.

Though we could not solve the problem entirely, we did apply two methods to increase the chance of finding the identifiers in OBX-5. The first method was to check whether identifiers present in the other 48 fields were also present in OBX-5, and then masked those identifiers and their variations. Variations were words that had the same meaning but were expressed in different ways. For example, a date could be spelled as "01-01-2022", "20220101", "Jan 01, 2022" and so on. This method could target most identifiers in OBX-5, except for identifiers not presented in the other 48 fields. We then developed a function to the program that allows users to add extra identifiers, which could possibly appear in OBX-5 to a "mask list" and increase the capture rate. Similarly, users could also add identifiers that they didn't want the program to mask to a "no mask list."

### RESULTS

Manual de-identification usually takes 10-30 minutes to process one HL7 message, while automatic de-identification, using this tool, takes only about 11 seconds to process 10,000 messages with an average file size. The tool accurately pinpoints the identifiers for 98% of data fields to be de-identified and the remaining 2% (OBX-5) are manually processed. Adding more identifiers to the "mask list" and "no mask list" increases the capture rate of identifiers in OBX-5 significantly. Furthermore, this rate would keep increasing as more ePath reports are received.

**TABLE 1: THE PERFORMANCE OF THE PROGRAM (IN SECONDS)**

| Number of Messages | Runtime 1 | Runtime 2 | Runtime 3 | Runtime 4 | Runtime 5 | Average |
|---|---|---|---|---|---|---|
| 10 | 0.046 | 0.046 | 0.047 | 0.047 | 0.048 | 0.047 |
| 100 | 0.238 | 0.235 | 0.226 | 0.248 | 0.249 | 0.239 |
| 1,000 | 1.454 | 1.437 | 1.407 | 1.448 | 1.453 | 1.440 |
| 10,000 | 10.987 | 11.169 | 10.910 | 10.939 | 10.968 | 10.994 |

### SUSTAINING SUCCESS

The de-identification tool has greatly increased efficiency and was applied to process ePath reports from other STAR participating states. More improvements are under development, including building a user interface and employing Natural Language Processing (NLP) models.

### STORY QUOTE

"The tool improves efficiency in sharing pathology data and allows the project to expand the development dataset as gaps are identified." – STAR Project

### REGISTRY CONTACT INFORMATION

DHHS.NebraskaCancerRegistry@nebraska.gov
Nebraska Cancer Registry Website

U.S. Department of Health and Human Services
Centers for Disease Control and Prevention