

# Data Quality Improvements in National Syndromic Surveillance Program (NSSP) Data

**Girum S. Ejigu\*, Lakshmi Radhakrishnan, Paul McMurray and Roseanne English**

Division of Health Informatics and Surveillance, Center for Surveillance, Epidemiology, and Laboratory Services, Centers for Disease Control and Prevention (CDC), Atlanta, GA, USA

## Objective

Review the impact of applying regular data quality checks to assess completeness of core data elements that support syndromic surveillance.

## Introduction

The National Syndromic Surveillance Program (NSSP) is a community focused collaboration among federal, state, and local public health agencies and partners for timely exchange of syndromic data. These data, captured in nearly real time, are intended to improve the nation's situational awareness and responsiveness to hazardous events and disease outbreaks. During CDC's previous implementation of a syndromic surveillance system (BioSense 2), there was a reported lack of transparency and sharing of information on the data processing applied to data feeds, encumbering the identification and resolution of data quality issues. The BioSense Governance Group Data Quality Workgroup paved the way to rethink surveillance data flow and quality. Their work and collaboration with state and local partners led to NSSP redesigning the program's data flow. The new data flow provided a ripe opportunity for NSSP analysts to study the data landscape (e.g., capturing of HL7 messages and core data elements), assess end-to-end data flow, and make adjustments to ensure all data being reported were processed, stored, and made accessible to the user community. In addition, NSSP extensively documented the new data flow, providing the transparency the community needed to better understand the disposition of facility data. Even with a new and improved data flow, data quality issues that were issues in the past, but went unreported, remained issues in the new data. However, these issues were now identified. The newly designed data flow provided opportunities to report and act on issues found in the data unlike previous versions. Therefore, an important component of the NSSP data flow was the implementation of regularly scheduled standard data quality checks, and release of standard data quality reports summarizing data quality findings.

## Methods

NSSP data was assessed for the national-level completeness of chief complaint and discharge diagnosis data. Completeness is the rate of non-null values (Batini et al., 2009). It was defined as the percent of visits (e.g., emergency department, urgent care center) with a non-null value found among the one or more records associated with the visit. National completeness rates for visits in 2016 were compared with completeness rates of visits in 2017 (a partial year including visits through August 2017). In addition, facility-level progress was quantified after scoring each facility based on the percent completeness change between 2016 and 2017. Legacy data processed prior to introducing the new NSSP data flow were not included in this assessment.

## Results

Nationally, the percent completeness of chief complaint for visits in 2016 was 82.06% (N=58,192,721), and the percent completeness of chief complaint for visits in 2017 was 87.15% (N=80,603,991). Of the 2,646 facilities that sent visits data in 2016 and 2017, 114 (4.31%)

facilities showed an increase of at least 10% in chief complaint completeness in 2017 compared with 2016. As for discharge diagnosis, national results showed the percent completeness of discharge diagnosis for 2016 visits was 50.83% (N=36,048,334), and the percent completeness of discharge diagnosis for 2017 was 59.23% (N=54,776,310). Of the 2,646 facilities that sent data for visits in 2016 and 2017, 306 (11.56%) facilities showed more than a 10% increase in percent completeness of discharge diagnosis in 2017 compared with 2016.

## Conclusions

The newly designed NSSP data flow provided more opportunity to identify data quality issues. By applying data quality checks within the newly designed NSSP data flow, data quality issues related to HL7 messages and processed data could be identified early. Improvements in data quality were demonstrated by measuring percent completeness of chief complaint and discharge diagnosis data in 2017 and comparing with data from 2016. Overall, several factors helped improve data quality: implementation of routine and targeted data quality checks; investigation of the root cause of data quality issues; and communication of such findings by engaging the NSSP team, sites, and vendors.

## Keywords

NSSP; Data Quality; Completeness; Chief Complaint; Discharge Diagnosis

## Acknowledgments

We thank Paula Yoon, David Walker, Michael Coletta, Alan Davis, Niketta Womack, NSSP Partners and BioSense Governance Group.

## References

- Batini, C., Cappiello, C., Francalanci, C. and Maurino, A. (2009) Methodologies for data quality assessment and improvement. *ACM Comput. Surv.*, 41(3). 1-52.

\*Girum S. Ejigu

E-mail: kwa7@cdc.gov

