September 2020

IMPROVING THE QUALITY AND COMPLETENESS OF ELECTRONIC HEALTH RECORD DATA USED IN SYNDROMIC SURVEILLANCE

Final Report

Acknowledgements

This report describes the work performed on the Improving the Quality and Completeness of EHR Data Used in Syndromic Surveillance project for the Council of State and Territorial Epidemiologists (CSTE) by Thought Bridge, LLC, under the direction of Barbara L. Massoudi, MPH, PhD. The Thought Bridge team led data collection, analysis and solution design activities described in this report during the project period of performance, October 1, 2019 – July 31, 2020. This work was performed under CSTE and Centers for Disease Control and Prevention (CDC) – Cooperative Agreement number 1 NU38OT000297-01-00. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of CDC.

CSTE and Thought Bridge would like to thank the following organizations for their participation in the EHR for Syndromic Surveillance Project Workgroup:

Alabama Department of Public Health
Alaska Department of Health and Social Services
Arizona Department of Health Services
Boston Public Health Commission
Centers for Disease Control and Prevention
Cerner Corporation
Connecticut Department of Public Health
Florida Department of Health
ICF International (on contract to CDC)
Illinois Department of Public Health
Kansas Department of Health and Environment
Massachusetts Department of Public Health
New Hampshire Department of Health and Human Services
New York State Department of Health
Ohio Department of Health
Tennessee Department of Health
Utah Department of Health
Virginia Department of Health
Washington State Department of Health
Wisconsin Department of Health Services

All key informant interviews were conducted by Dr. Barbara Massoudi with notetaking support by Ms. Diana Sobolevskaia, Research Assistant. Working group meetings were facilitated by Dr. Barbara Massoudi, with Ms. Diana Sobolevskaia providing notetaking support.

Table of Contents

Executive Summary

Electronic health records (EHRs) improve public and population health outcomes by efficiently collecting health information to be shared across health care organizations and public health agencies [1]. Unfortunately, there is great variability among EHR systems in data accuracy and completeness, data element definitions, and standards, which results in a lack of interoperability and data harmonization across systems. The purpose of this report is to summarize identified areas for EHR data quality improvement for syndromic surveillance, and to provide recommendations on how state, territorial, local, and tribal health departments and EHR vendor communities can collaborate to address.

The team took an approach which focused on determining the key areas to improve the quality of data received from EHRs for the purposes of syndromic surveillance. The team also proposed solutions to address these data quality issues and developed an operational plan for who would enact these solutions and the timeframe for their implementation.

Methods used included a literature review of publications related to EHR data quality (found in Appendix A), semi-structured key informant interviews (N=14), formation of an EHR Data Quality for Syndromic Surveillance Project Workgroup (N=27), and, six facilitated virtual discussions with the convened workgroup.

According to key informant interview, the most critical EHR data elements for syndromic surveillance were determined including: chief complaint, diagnosis codes, patient demographics (i.e., age, birthdate, sex, gender, race, ethnicity, home or work address, zip code), dates (i.e., admission, encounter, service, arrival), and reason for visit. Additionally, the critical aspects of data quality identified are completeness/presence of data element(s), timeliness, correctness/validity/accuracy, and comparability/reliability. Twenty-one individual data quality issues were identified via key informant interviews and workgroup discussions, which are summarized in four overarching categories: 1) Technical Guidance and Certification; 2) Changes, Updates, Customization, and Standards; 3) Processes and Workflow; and 4) Relationships. Identified issues mirror the often-cited informatics triad of *people, process*, and *technology*, where the first two elements are recognized as being the most challenging.

Twelve potential solutions to address identified issues were proposed and activities were determined by the workgroup including timeframe and who would lead these efforts. The recommendations are divided into short-term (six months or less) and long-term (greater than six months) depending on the timeframe needed for implementation. A brief bulleted list of the recommendations is presented below.

- Clarify current PHIN Implementation Guide on CDC NSSP website
- Archive outdated guides on CDC NSSP website
- Update/Revise CDC NSSP website text to indicate current Guide (v.2.0) is certifying to 2015 [2]
- Determine the feasibility of conducting a review of the current flexibility built into the PHIN Implementation Guide with the aim of achieving a balance between flexibility and data quality
- Conduct review of HL7-balloted Messaging Guide, noting any changes/updates/corrections needed and considering addition of new fields needed to address COVID-19
- Implement enhanced training for sites on using the PHIN Implementation Guide and the HL7 Messaging Guide in tandem, such as a video orientation
- Develop NIST tool tutorial for site to use with facilities
- Update and address issues with the NIST validation tool

- Add the Data Quality Dashboard tools to the ‘Staging Phase’ of onboarding
- Share the Data Quality Tools on Demand SAS Program with the ESSENCE Community
- Provide access to all data quality tools from staging to production phases
- Plan a Syndromic Surveillance Data Quality Workshop involving sites, vendors, facilities and CDC

Background

The implementation of electronic health records (EHR) throughout the U.S. health care system presents new opportunities for enhancing and expanding data use for public health surveillance. EHRs can include a patient's medical history, diagnoses, medications, treatment plans, immunization dates, allergies, radiology images, and/or laboratory and test results. EHRs may also provide data on populations, geographic areas, and health conditions that are not fully captured by traditional surveillance mechanisms such as population-based surveys or reportable disease systems. The adoption of EHRs and use of EHR data has been promoted through public health programs, such as Meaningful Use and the Promoting Interoperability Program, which incentivize public health reporting for syndromic surveillance purposes.

Electronic health records (EHRs) can improve public and population health outcomes by efficiently collecting electronic health information that can be shared across health care organizations and to public health agencies [1]. Specifically, EHRs can improve public health reporting and surveillance by automating data feeds to syndromic surveillance systems, improving the timeliness and accuracy of those reports [3]. Leveraging EHR data for these public health activities has the potential to decrease the reporting burden on healthcare partners and could promote more timely, accurate, and complete data sharing between healthcare and public health. Unfortunately, there is great variability among EHR systems in data accuracy and completeness, and data element definitions, and standards, which results in a lack of interoperability and data harmonization across systems. Poor or unknown data quality and a lack of interoperability threaten the benefits that EHR data could provide to both healthcare and public health.

The purpose of this report is to summarize identified areas for EHR data quality improvement for syndromic surveillance, and to provide recommendations on how state, territorial, local, and tribal health departments and EHR vendor communities can collaborate to address them.

Methods

In order to improve the quality and completeness of EHR data used in syndromic surveillance a literature review was conducted first. Next, a series of interviews were conducted with key informants to collect information from state and local health professionals. Following the interviews, a workgroup was established to address the common themes found within the interviews.

Literature Review

Methods used included a literature review of publications related to EHR data quality. The literature review and full description of methods used in preparing it can be found in Appendix A.

Key Informant Interviews

Semi-structured key informant interviews were the primary method for collecting information about 1) key areas to improve the quality of data received from EHRs for the purposes of syndromic surveillance 2) identified data quality issues with EHR data, and 3) proposed solutions to address described data quality issues. Discussions were guided by a series of questions that took place during January through March 2020 (see Appendix B).

Key informant interviews were designed to collect information from state and local public health professionals, EHR vendors, health information exchanges, academic public health experts, government agencies, and healthcare organizations about topics related to the quality of EHR data used for syndromic surveillance. These topics included:

- Identification of EHR data elements considered to be most critical for efficient and effective syndromic surveillance use
- Identification of the most important data quality considerations when using EHRs for syndromic surveillance
- Awareness of formal assessments or evaluations of data quality of EHR data used for syndromic surveillance
- Feedback on findings of the literature review
- Awareness of other resources suggested for review
- Recommendations of other experts to be included in the key informant interviews
- Recommendations to improve identified data quality issues or address problems with data quality

Twenty individuals actively involved in syndromic surveillance and having expertise with EHR data used for syndromic surveillance were invited to participate as key informants. The state and local public health agency key informants were identified based on their participation on the National Syndromic Surveillance Program's (NSSP) Community of Practice (CoP) Data Quality Subcommittee (DQ), which engages members to identify and address syndromic data quality issues and concerns through thoughtful discussion and inclusion of outside stakeholders [4]. EHR vendors and health information exchanges were invited to participate as they actively submit data to NSSP. Invited academic public health key informants have published on the use of EHRs in syndromic surveillance, and federal health agency key informants provided the national public health perspective. Organizational affiliations of key informants who participated in interviews are illustrated in Table 1. Although invited, no healthcare organizations were able to participate in the key informant interviews.

Table 1. Organizational affiliations of key informants.

Key Informant Organization Type	Number of Key Informant (N=14)
State and Local Public Health Agencies	4
EHR Vendors	2
Health Information Exchanges	2
Academic Public Health	2
Federal Health Agencies	4
Healthcare Organizations	0

EHR for Syndromic Surveillance Workgroup Facilitated Discussions

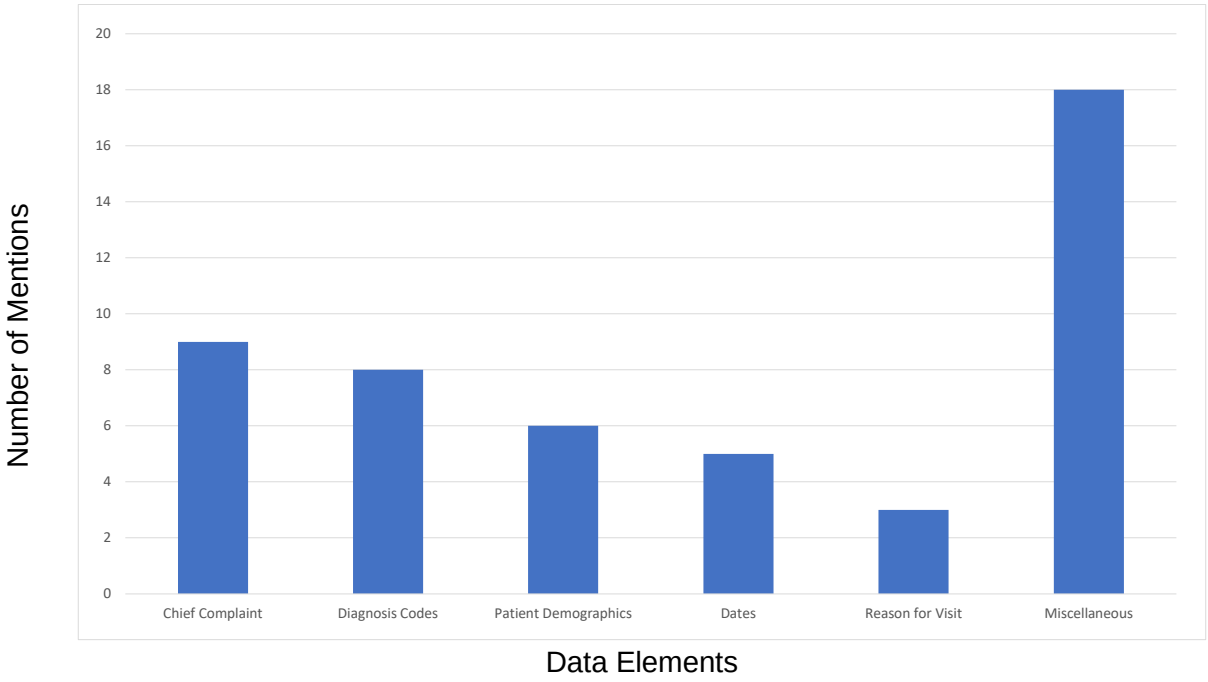
CSTE solicited members of the NSSP CoP DQS to convene a project-specific workgroup. The EHR for Syndromic Surveillance Project Workgroup consisted of 27 members who expressed interest in participating, and the workgroup met monthly between February and July 2020 (total of six meetings). Workgroup meetings allowed for detailed discussions of the project approach, review of literature review and key informant findings, developing recommendations and consensus building. Meetings included project progress updates and facilitated discussion questions, which were prepared and distributed to members in advance. Workgroup calls were recorded and recordings were available afterward for points of clarification and review of notes. Meeting agendas and number of participants for each meeting are listed in Appendix C.

Results

Key Data Elements Identified for Improving the Quality of EHR data

Figure 1 demonstrates the EHR data elements determined to be key in improving the quality of data received from EHRs for the purposes of syndromic surveillance via the key informant interviews process. In descending order, the key data elements are chief complaint, diagnosis codes, patient demographics (i.e., age, birthdate, sex, gender, race, ethnicity, home or work address, zip code), dates (i.e., admission, encounter, service, arrival), reason for visit, and other data elements (classified as miscellaneous). The miscellaneous category (and number of mentions) includes triage notes (2), problem list (1), procedure codes (2), discharge disposition (2), facility name (2), facility address (2), arrival date/time (1), patient class (1), travel history (1), clinical impression (1), transitions of care (1), primary care medical home (1), and medical record number (1).

Figure 1. Key Data Elements for Improving the Qualitiy of EHR Data by Number of Mentons.



Most Important EHR Data Quality Considerations for Syndromic Surveillance

The critical aspects of data quality identified through the key informant interviews in descending order are completeness/presence of data element(s), timeliness, correctness/validity/accuracy and comparability/reliability.

Issues of EHR Data Quality Identified for Syndromic Surveilllance

Data quality issues identified through the key informant interviews and the EHR for Syndromic Surveillance Project Workgroup meetings were numerous and spanned many areas. For ease of presentation and to facilitate development of holistic solutions, the issues were grouped into four overarching categories: 1) Technical Guidance and Certification; 2) Changes, Updates, Customization, and Standards; 3) Processes and Workflow; and 4) Relationships. Table 2 demonstrates the individual data quality issues gathered, organized within the four groups above.

Table 2. Data quality issues, grouped by four categories.

1. Technical Guidance and Certification
1.1 Issues related to the PHIN Implementation Guide
Multiple versions of the Guide are listed on CDC NSSP's website making it unclear to readers which version is current
Guide is not well developed, and not of the same rigor level as the HL7 messaging guide[1]
Guide lacks EHR vendor agreement/buy-in
Guide was developed in a void and without the necessary NSSP Community of Practice participation
Guide specifications are often not followed (e. g., data type disagreements, facility type coding)
Guide lacks enforcement and clear identification of who should enforce it
1.2 Issues related to HL7 Certification
Systems may be certified but are not able to produce certified messages
2. Changes, Updates, Customization, and Standards
Jurisdictions frequently request customizations
No process is in place to characterize and prioritize requested changes as customization versus future development, and urgent versus routine.
No risk/impact assessment is required before changes are implemented
No standards are imposed across jurisdictions related to changes and customization
Proposed new data elements are not consistently evaluated for inclusion
3. Processes and Workflow
Processes and tools for monitoring data quality are not shared/utilized consistently nor widely
Information contained in A01 (admit) messages conflicts with information coming in A08 (update) messages
Different identification numbers are assigned across different hospital departments
Chief complaint and reason for visit fields are inconsistently applied across facilities
Multiple issues need to be overcome by hospitals so that data gets captured in the workflow
4. Relationships
Some jurisdictions lack collaborative relationships with their data providers and vendors

[1] "HL7 Version 2.5.1 Implementation Guide: Syndromic Surveillance, R1 – US Realm", Syndromic Surveillance Message Guide.

Some jurisdictions look to CDC to resolve their issues with data providers and vendors
Organizational silos prevent resolution of data quality issues
Communication between jurisdictions, data providers, and vendors can be difficult

Recommendations

Based on workgroup discussions, potential recommendations to address the identified data quality issues were developed. Solutions were divided into short-term and long-term, depending on the timeframe needed for implementation. Short-term solutions are straightforward and can be completed in six months or less. Long-term solutions are either more complex or necessitate building partnerships and will require more than six months to implement[2]. Appendix D contains a table with the proposed solutions, lead organization(s), and timeframes for implementation to address the identified data quality issues.

Proposed Short-Term Solutions to Identified Data Quality Issues

Short-term issues primarily include those designed to address Technical Guidance. It is feasible that these solutions can be implemented within the next six months. A description of the solution, issues addressed, activities, timeframe and lead organization(s) are outlined below.

1. **Clarify current PHIN Implementation Guide on CDC NSSP website**
   Identified location to address solution: https://www.cdc.gov/nssp/technical-pubs-and-standards.html#Guidance
   This recommendation addresses the Technical Guidance data quality issue that "*multiple versions of the Guide are listed on CDC NSSP's website, making it unclear to readers which version is current.*" This solution requires a critical review of the many documents listed in the table under the section heading Message Mapping Guides listed on the website, and designing a way to clearly indicate which is the current version (v.2.0) and is estimated to take approximately two months to complete. This activity should be led by NSSP.

2. **Archive outdated Guides on CDC NSSP website**
   Identified location to address solution: https://www.cdc.gov/nssp/technical-pubs-and-standards.html#Guidance
   This recommendation addresses the Technical Guidance data quality issue that "*multiple versions of the Guide are listed on CDC NSSP's website, making it unclear to readers which version is current.*" This solution requires a critical review of the many documents listed in the table under the section heading Message Mapping Guides and creating an archive tab on the NSSP website to access older versions of guide. This effort is estimated to take approximately two months to complete. This activity should be led by NSSP.

3. **Update/revise CDC NSSP website text to indicate current Guide (v.2.0) is certifying to 2015 [2].**
   Identified location to address solution: https://www.cdc.gov/nssp/technical-pubs-and-standards.html#Guidance
   This recommendation addresses the Technical Guidance data quality issue that "*multiple versions of the Guide are listed on CDC NSSP's website, making it unclear to readers which version is current.*" This solution requires a change to the table under the section heading Message Mapping Guides updating the text to read "Certifying 2015 Edition Health IT Modules" and is estimated to take approximately two months to complete. This activity should be led by NSSP.

[2] Participants indicated that this was partly due to the demands of the sustained public health COVID-19 response.

4. **Update and address issues with the NIST validation tool.**
   This recommendation addresses the Technical Guidance data quality issue that "*systems may be certified but are not able to produce certified messages.*" This solution requires a review of the validation tool to determine a complete list of existing issues and then making changes to address those. This solution implementation is estimated to take approximately 6 months to complete and should be led by the NSSP CoP Data Quality Subcommittee or a workgroup under the subcommittee, with substantial participation from the Office of the National Coordinator for Health Information Technology (ONC) and the National Institute of Standards and Technology (NIST).

Proposed Long-Term Solutions to Identified Data Quality Issues

Long-term issues include primarily those designed to address Technical Guidance, Processes and Workflow, and Data Quality Workshop. These solutions are expected to require more than six months to implement. A description of the solution, issues addressed, activities, timeframe and lead organization(s) are outlined below.

1. **Determine the feasibility of conducting a review of the current flexibility built into the PHIN Implementation Guide with the aim of achieving a balance between flexibility and data quality.**
   This recommendation addresses the Technical Guidance data quality issue that "*guide is not well developed and is not to same level or rigor as HL7 messaging guide.*" This solution requires a thorough review at the individual specification level of the current Guide (v.2.0) and a determination if the intentional flexibility originally designed for the Guide is still prudent, given the impacts on data quality and is estimated to take approximately nine months to complete. This activity should be led by the NSSP CoP Data Quality Subcommittee or a workgroup under the subcommittee.

2. **Conduct review of HL7-balloted Messaging Guide, noting any changes/updates/corrections needed and considering addition of new data elements needed to address COVID-19[3].**
   This recommendation addresses the Technical Guidance data quality issue that "*systems may be certified but are not able to produce certified messages.*" This solution requires a thorough review and modification of the HL7-balloted Messaging Guide including incorporation of the several already-identified changes, along with the incorporation of any new fields needed to address COVID-19. This solution implementation is estimated to take approximately nine months to complete and is dependent on the pilot to begin in October 2020. This effort should be led by NSSP with substantial participation from the NSSP CoP Data Quality Subcommittee, ONC, and Health Level Seven International.

3. **Implement enhanced training for sites on using the PHIN Implementation Guide and the HL7 Messaging Guide in tandem, such as a video orientation.**
   This recommendation addresses the Technical Guidance data quality issues that "*guide specifications are often not followed*" and "*systems may be certified but are not able to produce certified messages.*" This solution is aimed at increasing compliance with the Guides. This effort will require development of learning objectives and an enhanced training to fulfill those objectives, presenting both the PHIN and HL7 Guides together in an integrated manner. It is estimated that this activity will take approximately

[3] If the new COVID-19 data elements are already referenced in the HL7 base standard 2.5.1., changes to update the use from optional to required should occur in the SyS Implementation Guide. If they are not currently included in the standard, then a request would need to be made to update the 2.5.1. standard first.

four months to complete and should not be started until the updates to the PHIN and HL7 Guides are completed. A video orientation to the two Guides might be a preferred format for delivering this training. This activity should be led by the NSSP with the NSSP CoP Data Quality Subcommittee.

4. **Develop NIST tool tutorial for sites to use with facilities.**
   This recommendation addresses the Technical Guidance data quality issue that "*systems may be certified but are not able to produce certified messages.*" This solution requires establishing learning objectives and determining modes of delivery and content for the tutorial. Information to be included in the tutorial should include the rationale for use of the tool and how to read the reports generated after testing with the tool. This activity will take approximately six months but should be done after or in conjunction with the update to the validation tool and should be led by NSSP with the NSSP CoP Data Quality Subcommittee.

5. **Add the NSSP Data Quality (DQ) Dashboard tools to the "Staging Phase" of onboarding.**
   Identified location to address solution:
   https://dashboards.syndromicsurveillance.org/app/dqDashboard
   This recommendation addresses the Processes and Workflow data quality issue *"processes and tools for monitoring data quality are not consistently and widely shared/utilized."* This solution requires a policy review, changes in access to the tools, and updating of onboarding processes and materials. This activity will take approximately six to eight months and should be led by the NSSP with participation from members of the NSSP CoP Data Quality Subcommittee.

6. **Share the Staging Data Quality on Demand SAS Tool with the NSSP Community of Practice[4].**
   This recommendation addresses the Processes and Workflow data quality issue that "*processes and tools for monitoring data quality are not consistently and widely shared/utilized.*" This solution requires a policy review, changes in access to the tools, training and technical assistance support, and communication messaging. This activity will take approximately six to eight months and should be led by the NSSP with participation from members of the NSSP CoP Data Quality Subcommittee.

7. **Provide access to all data quality tools from staging to production phases.**
   This recommendation addresses the Processes and Workflow data quality issue that "*processes and tools for monitoring data quality are not consistently and widely shared/utilized.*" This solution requires a policy review, changes in access to the tools, training and technical assistance support, and communication messaging. This activity will take approximately six to eight months and should be led by the NSSP with participation from members of the NSSP CoP Data Quality Subcommittee.

8. **Plan a Syndromic Surveillance Data Quality Workshop involving sites, vendors, facilities, and CDC**.
   There are three proposed objectives/tracks for the workshop, each addressing different solution topics and solutions.

   Track I: Development of change request/enhancements process and guidelines.

[4] This site is password protected and requires a NSSP BioSense Platform Access & Management Center (AMC) password. Contact the NSSP Service Desk to request an account.

This recommendation addresses the Changes, Updates, Customization, and Standards data quality issues that "*jurisdictions frequently request customizations*"; "*no process is in place to characterize and prioritize requested changes as customization versus future development, and urgent versus routine*"; "*no risk/impact assessment is required before changes are implemented*"; "*no standards are imposed across jurisdictions related to changes and customization*", and "*proposed new data elements are not consistently evaluated for inclusion.*"

Track II: Development of case studies for successful collaborations, including details related to:

- Networking, coalition building, and strengthening relationships with vendors and facilities
- Coordinating with multiple facility partners – hospital Chief Information Officer (CIO), clinical/admit staff, and IT staff

This recommendation addresses the Relationships data quality issues that "*some jurisdictions lack collaborative relationships with their data providers and vendors*"; "*some jurisdictions look to CDC to resolve their issues with data providers and vendors";* "*organizational silos prevent resolution of data quality issues*"; and "*communication between jurisdictions, data providers, and vendors can be difficult.*"

Track III: Development of recommendations and best practices to understand, work with, and not disrupt facility workflow while collecting required data elements for syndromic surveillance.

This recommendation addresses the Processes and Workflow data quality issues that "*chief complaint and reason for visit fields are inconsistently applied across facilities*" and "*multiple issues need to be overcome by hospitals so that data gets captured in the workflow.*" Implementing this recommendation will require a group of individuals to plan workshop details including objectives, mode of delivery, location and costs of travel (including transportation, housing, meals, incidentals, and expenses if applicable), participants, agenda, facilitation, and participant/track activities. The planning and implementation of the workshop will take approximately nine to twelve months and should be led by a workgroup under the NSSP CoP Data Quality Subcommittee, with CDC and ONC co-sponsoring, and CSTE hosting.

Conclusion

The chief aim of syndromic surveillance is to obtain information quickly, providing the potential for immediate investigation and identification of emerging public health issues. Ultimately, this allows for earlier public health intervention which enables risk mitigation, and morbidity and mortality reduction. It is not surprising then, that the most important data elements identified for syndromic surveillance relate to the epidemiologic triad of *person, place*, and *time*. Key elements of *person* include patient demographic information to better characterize events and risks, and patient clinical information such as chief complaint, reason for visit, and diagnostic codes, as available. The most critical aspects of data quality were availability of these data elements, and timeliness, with correctness and comparability less important within the syndromic context.

Data quality issues identified fell into four categories: 1) Technical Guidance and Certification, 2) Changes, Updates, Customizations, and Standards, 3) Processes and Workflow, and 4) Relationships. These issues mirror the often-cited informatics triad of *people, process*, and *technology*, where the first two elements are recognized as being the most challenging. In this project, *people* issues are central to the relationship-related impacts on data quality, with *process* issues identified in both changes, updates, etc., and processes and workflow. Finally, *technology* has the least and probably most easily addressed issues associated with it, falling under the topic of technical guidance and certification. The proposed solutions reflect these varying levels of detail in both their approaches and timeframes for implementation. Technical issues with the PHIN and HL7 Guides can be addressed more easily and quickly with a fewer number of involved parties, while processes, workflow, changes, updates, and relationship issues will likely require a workshop with dedicated tracks and broad stakeholder representation to work through the complicated, and perhaps, charged sub-issues that resulted in these problems.

References

1. Friedman DJ, Parrish RG, Ross DA. Electronic health records and US public health: current realities and future promise. *Am J Public Health*. 2013;103(9):1560-1567. doi:10.2105/AJPH.2013.301220
2. Office of the National Coordinator for Health Information Technology (ONC), Department of Health and Human Services (HHS). 2015 Edition Health Information Technology (Health IT) Certification Criteria, 2015 Edition Base Electronic Health Record (EHR) Definition, and ONC Health IT Certification Program Modifications. Final rule. *Fed Regist*. 2015;80(200):62601-62759.
3. Gould DW, Walker D, Yoon PW. The Evolution of BioSense: Lessons Learned and Future Directions. *Public Health Rep*. 2017;132(1_suppl):7S-11S. doi:10.1177/0033354917706954.
4. NSSP Community of Practice. https://www.cste.org/page/nsspcop. Accessed July 15, 2020.

Appendices

A – Literature Review
B – Key Informant Questions
C – EHR for Syndromic Surveillance Workgroup Discussion Agendas and Participation
D – Table of Proposed Solutions, Leads, and Timeframes

Appendix A – Literature Review

Introduction

Electronic health records (EHRs) can improve public and population health outcomes by efficiently collecting electronic health information that can be shared across health care organizations and to public health agencies, and then leveraged to improve the health of the public. Specifically, EHRs can improve public health reporting and surveillance through syndromic surveillance by automating data feeds to syndromic surveillance systems, improving the timeliness and accuracy of those reports, while minimizing resources needed to onboard individual health care organizations. The usage and functionality of EHRs have increased throughout the past decade. Researchers and public health organizations have used the syndromic surveillance data collected from EHRs to conduct more complex epidemiologic investigations (Casey, 2016). As the quantity and quality of data available to public health organizations increases, they can better monitor, prevent, and manage disease. The focus of this literature review is on data quality of EHR data for syndromic surveillance and is part of a larger effort to identify ways to improve collaboration between STLT and EHR vendor communities.

Methods

For this review, we conducted searches of the biomedical informatics literature published in MEDLINE. We used the following MEDLINE Subject Headings (MeSH) keyword searches to maximize sensitivity: “electronic health record” AND “completeness”, and “electronic health record” AND “data accuracy”.

We used inclusion and exclusion factors to narrow the list of articles. To be included an article had to focus on a topic related to health care or public health systems, be published in English, and focused on data quality as the primary subject of the study or a main component of the study methodology. Original research articles designed to assess the data quality of EHRs along with articles detailing the frameworks and methodologies applied to these studies, and recommendations for improvement to EHR data quality were included. Editorials were also included in the full review to gain additional perspective on the issues.

Each article was reviewed with an eye for specific data elements within EHRs that were assessed for data quality. The author determined which dimension(s) of data quality were measured, and by which method.

Results

The initial search results from MEDLINE produced 475 citations. After reviewing the titles of these, 53 were selected for full abstract review. This resulted in 38 citations included in the final full review. All articles included in the final full review are listed in the references. Five additional articles were identified through review of these 38 and were included in the final group, resulting in a total of 43 full reviews.

From the selected papers, we developed the following framework of domains, which are listed in Table 1, along with their descriptions, and synonyms based on the works of Weiskopf (2013b) and Johnson (2015). Weiskopf derived five dimensions of data quality based on a review of the literature. These dimensions were defined as: completeness, correctness, concordance, plausibility, and currency. Johnson (2015), built on these to develop a data quality ontology using UML (Unified Modeling Language) to represent high-level core data quality dimensions, which included: correctness, consistency, completeness, and currency. Dimensions of data quality discussed in all other reviewed publications (Brown, 2016; Feder, 2018; Kilkenny, 2018; Raman, 2018; Terry, 2019) are represented in this framework.

Table 1. Terms, definitions, and synonyms used in the literature to describe the four common dimensions of data quality.[3]

Term	Completeness	Correctness	Comparability	Currency
Definition	The proportion of observations made about the world that were recorded in the [EHR].	The proportion of [EHR] observations that are a correct representation of the true state of the world.	The degree to which [EHR] data are consistent with, or comparable to, an external data source.	Is an element in the EHR a relevant representation of the patient state at a given point in time?
Synonyms	*Accessibility* *Availability* *Missingness* *Omission* *Presence* *Quality* *Rate of recording* *Sensitivity*	*Accuracy* *Credibility* *Corrections made* *Errors* *Misleading* *Positive predictive value* *Quality* *Validity*	*Agreement* *Concordance* *Consistency* *Reliability* *Variation*	*Recency* *Timeliness*

Because most articles have assessed the data quality domain of completeness, Weiskopf (2013a) further describes four different dimensions of the definition of completeness: documentation, breadth, density, and predictive [value]. A high level of documentation is achieved when a patient has a complete record. A high level of breadth is achieved when a patient has a record with all the required elements present. A high level of density is achieved when a patient has a record with multiple observations of the same measure over time. A high predictive level is achieved when a record includes a sufficient amount of information to predict a phenomenon of interest.

Based on her review of the literature, Weiskopf (2013b) described the most common methods of data quality assessment, which fell into seven broad categories, and are described below.

1. Gold standard – a dataset drawn from another source or multiple sources, with or without data from the EHR, which is used as a gold standard.
2. Data element agreement – two or more elements within an EHR are compared to see if they report the same or compatible information.
3. Element presence – a determination is made as to whether or not desired or expected data elements are present.
4. Data source agreement – data from the EHR are compared with data from another source to determine if they are in agreement.
5. Distribution comparison – distributions or summary statistics of aggregated data from the EHR are compared with the expected distributions for the clinical concepts of interest. *Note:* when considering the data quality of individual patient EHRs, this measure is not particularly helpful as it utilizes large groups of patients for comparison.
6. Validity check – data in the EHR are assessed using various techniques that determine if values "make sense".
7. Log review – information on the actual data entry practices (e.g. dates, times, edits) is examined.

Table 2 shows the results of the literature review by domain and data element. For each data element within a particular domain, the findings from the literature are listed, along with the citations from which those findings were drawn and the method(s) of assessment used.

Table 2. Data Elements by Domains of Data Quality.

Data Elements by Domain	Completeness	Correctness	Comparability	Currency
Patient demographic information				
Allen-Graham, 2018 Element presence	High		High	
Race/ethnicity				
Lee, 2016 Element presence Data source agreement	Low		Low	
Adverse drug events				
Allen-Graham, 2018 Element presence	Very low			
Paul & Robinson, 2012 Review article	Low			
Family history				

Tu, 2015 Element presence	Medium			
Allergies				
Allen-Graham, 2018 Element presence	Medium-High		Medium	
Paul & Robinson, 2012 Review article	Low			
Tu, 2015 Element presence	High			
Terry, 2019 Distribution comparison			Low-Medium	
Medical history/problem list				
Allen-Graham, 2018 Element presence	Medium		Medium	
Davey, 2013 Data source agreement	Low			
Tu, 2015 Element presence	High			
Wright, 2015 Element presence	Medium-High			
Terry, 2019 Distribution comparison			Medium-High	
Diagnoses				
Terry, 2019 Distribution comparison (diabetes) (hypertension) (hypothyroidism) (asthma) (obesity) (urinary tract infection) (pregnancy)	 High Medium Medium Medium Low High	 High Medium Medium Low Medium Low	 High High High Low High N/A	 High Medium

Surgical history				
Allen-Graham, 2018 Element presence	Medium		Medium	
Cause of death				
Haghighi, 2013 Gold standard		Low		
Treatment				
Hanafi, 2012 Element presence	Low			
Tu, 2015 Element presence	High			
Visit documentation				
Tu, 2015 Element presence	High			
Prescriptions				
Tu, 2015 Element presence	High			
Laboratory tests				
Tu, 2015 Element presence	Medium			
Consultation letters/Referrals				
Tu, 2015 Element presence	Low			
Height/Weight				
Tu, 2015 Element presence	Low			
Height				
Terry, 2019 Distribution comparison	Medium			Low

Weight				
Terry, 2019 Distribution comparison	Medium			Medium
Blood pressure				
Tu, 2015 Element presence	Low			
Terry, 2019 Distribution comparison	High			High
Immunizations				
Tu, 2015 Element presence	High			
Risk factors				
Tu, 2015 Element presence	Low			
Personal traits[1]				
Tu, 2015 Element presence	High			

[1] The author's term for sex, age, neighborhood income quintile, and rurality.

Discussion

Completeness was the dimension of data quality examined in most of the reviewed citations, followed by comparability. Few studies looked at the dimensions of correctness and currency. Within completeness there was a great deal of variability in the findings, with variations from low to high seen across studies and using different methods. Within the diagnosis data element, different ratings of data quality were noted for different conditions. Data elements of particular interest to public health, including risk factors, cause of death, race/ethnicity, blood pressure, family history, immunization, personal traits, height/weight, and patient demographic information showed a great deal of variation with most being of low to medium quality, with the exceptions of personal traits, patient demographic information, and immunizations. Data elements that are supplied to the EHR through interface or interoperability with other health information technology systems (immunizations, prescriptions, laboratory tests, patient demographic information, and visit documentation), were generally of higher quality than those which required manual data entry by healthcare providers.

For the most part, comparability showed the same general patterns as completeness. Correctness was only assessed for diagnoses and cause of death, and was found to vary depending on the particular condition being diagnosed. Currency was measured only for asthma, pregnancy, height, weight, and blood pressure and quality varied depending on which data element was being considered.

Data elements of particular interest for syndromic surveillance, other than diagnosis, were not measured in these studies, including chief complaint in emergency department records, patient admission, and facility name/identifier.

Conclusion

EHR data quality varies widely across the particular data elements and dimensions of quality studied. Data being provided through interface or interoperability with other electronic health information systems is of higher quality. Data of particular interest to public health, and specifically for syndromic surveillance was of lower quality, or has not been assessed or published in the peer-reviewed literature.

Literature Review References

Agniel D, Kohane IS, Weber GM. (2018). Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. British Medical Journal. 361:k1479. Erratum in British Medical Journal. 363:k4416.

Allen-Graham J, Mitchell L, Heriot N. (2018). Electronic health records and online medical records: an asset or a liability under current conditions? Australian Health Review. 42(1):59-65.

Birkhead GS(1), Klompas M, Shah NR. (2015). Uses of electronic health records for public health surveillance to advance public health. Annual Review of Public Health. 36:345-59.

Brown ML. (2016). Can't you just pull the data? The limitations of using of the electronic medical record for research. Paediatric Anaesthesiology. 26(11):1034-1035.

Callen J. (2014). What is the impact of electronic health records on the quality of health data? Health Information Management. 43(1):42-3.

Casey J A, Schwartz B S, Stewart W F, Adler N E. (2016). Using Electronic Health Records for Population Health Research: A Review of Methods and Applications. Annual Review Public Health 2016. 37:61–81.

Chan KS, Fowles JB, Weiner JP. (2010). Review: electronic health records and the reliability and validity of quality measures: a review of the literature. Medical Care Research Review. 67(5):503-27.

Charnock V. (2019). Electronic healthcare records and data quality. Health Informatics Library Journal. 36(1):91-95.

Davey, M., Sloan, M., Palma, S., Riley, M. and King, J. (2013). Methodological processes in validating and tilizati the quality of population-based data: a case study using the Victorian Perinatal Data Collection. Health Information Management Journal 42(3): 12-19.

De Lusignan S. (2016). In this issue – Don't make assumptions about integrated systems, data quality, tilization of technology, or access to routine data. Journal of Innovation in Health Informatics. 23(3):911.

Eden KB, Totten AM, Kassakian SZ et al. (2016). Barriers and facilitators to exchanging health information: a systematic review. International Journal of Medical Informatics. 88:44-51.

Feder SL. (2018). Data Quality in Electronic Health Records Research: Quality Domains and Assessment Methods. West Journal of Nursing Research. 40(5):753-766.

Freestone, D., Williamson, D. and Wollersheim, D. (2012). Geocoding coronial data: tools and techniques to improve data quality. Health Information Management Journal 41(3): 4-12.

Haghighi, M.H.H., Dehghani, M., Teshizi, S. H. and Mahmoodi, H. (2013). Impact of documentation errors on accuracy of cause of death coding in an educational hospital in Southern Iran. Health Information Management Journal. doi:10.12826/18333575.2013.0015.Haghighi

Hanafi, S., Hayatshahi, A., Torkamandi, H. and Javadi, M.R. (2012). Evaluation of treatment with Oseltamivir during the 2009 H1N1 (swine flu) pandemic: the problem of incomplete clinical information. Health Information Management Journal 41(1): 31-34.

Johnson SG, Speedie S, Simon G, Kumar V, Westra BL. (2015). A data quality ontology for the secondary use of EHR data. AMIA Annual Symposium Proceedings. 1937-46.

Johnson SG, Speedie S, Simon G, Kumar V, Westra BL. (2016). Application of an ontology for characterizing data quality for a secondary use of EHR data. Applied Clinical Informatics. 7(1):69-88.

Kilkenny MF, Robinson KM. (2018). Data quality: "Garbage in – garbage out". Health Information Management. 47(3):103-105.

Lee SJ, Grobe JE, Tiro JA. (2016). Assessing race and ethnicity data quality across cancer registries and EMRs in two hospitals. Journal of the American Medical Informatics Association. 23(3):627-34.

Massoudi BL, Marcial LH, Tant E et al. (2016). Using health information exchanges to calculate clinical quality measures: A study of barriers and facilitators. Healthcare. 4(2):104-8.

McConnell RA, Kane SV. 2018. The Potential and Pitfalls of Using the Electronic Health Record to Measure Quality. American Journal of Gastroenterology. 113(8):1111-1113.

Paul, L. and Robinson, K.M. (2012). Capture and documentation of coded data on Adverse Drug Reactions: an overview. Health Information Management Journal 41(3): 27-36.

Perry WM, Hossain R, Taylor RA. (2018). Assessment of the feasibility of automated, real-time clinical decision support in the emergency department using electronic health record data. BMC Emergency Medicine. 18(1):19.

Polnaszek B, Gilmore-Bykovskyi A, Hovanes M. (2016). Overcoming the challenges of unstructured data in multisite, electronic medical record-based abstraction. Medical Care. 54(10):e65-72.

Raman SR, Curtis LH, Temple R et al. (2018). Leveraging electronic health records for clinical research. American Heart Journal. 202:13-19.

Rowlands. S., Callen, J. and Westbrook, J. (2012). Are general practitioners getting the information they need from hospitals to manage their lung cancer patients? A qualitative exploration. Health Information Management Journal 41(2):4-13.

Schiff, G.D. and Bates, D.W. (2010). Can electronic clinical documentation help prevent diagnostic errors? New England Journal of Medicine 362(12): 1066-1069.

Soto M, Capurro D, Catalán S. (2015). Evaluating the data completeness in the Electronic Health Record after the Implementation of an Outpatient Electronic Health Record. Studies in Health Technology and Informatics. 216:885.

Ta CN, Weng C. (2019). Detecting systemic data quality issues in electronic health records. Studies in Health Technology and Informatics. 264:383-387.

Terry AL, Stewart M, Cejic S et al. (2019). A basic model for assessing primary health care electronic medical record data quality. BMC Medical Informatics and Decision Making. 12;19(1):30.

Tu K, Widdifield J, Young J et al. (2015). Are family physicians comprehensively using electronic medical records such that the data can be used for secondary purposes? A Canadian perspective. BMC Medical Informatics and Decision Making. 15:67.

Tubaishat A. (2019). The effect of electronic health records on patient safety: A qualitative exploratory study. Informatics for Health and Social Care. 44(1):79-91.

Vuokko R, Mäkelä-Bengs P, Hyppönen H et al. (2015). Secondary use of structured patient data: interim results of a systematic review. Studies in Health Technology and Informatics. 210:291-5.

Walji MF. (2019). Electronic health records and data quality. Journal of Dental Education. 2019 83(3):263-264.

Weiskopf NG, Hripcsak G, Swaminathan S, et al. (2013a). Defining and measuring completeness of electronic health records for secondary use. Journal of Biomedical Informatics 46(5):830-6.

Weiskopf NG, Weng C. (2013b). Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. Journal of the American Medical Informatics Association. 20(1):144-51.

Weng CY. (2017). Data accuracy in electronic medical record documentation. JAMA Ophthalmology. 135(3):232-233.

Wright A, McCoy AB, Hickman TT, et al. (2015). Problem list completeness in electronic health records: A multi-site study and assessment of success factors. International Journal of Medical Informatics. 84(10):784-90.

Appendix B – Key Informant Questions

Improving the Quality and Completeness of EHR Data Used in Syndromic Surveillance

Key Informant Interview Questionnaire:

1. Please tell us what your organization does in syndromic surveillance and your role.
2. What data elements do you consider to be most important for syndromic surveillance?
3. What do you think are the most important aspects of data quality for EHRs used in syndromic surveillance?
4. Are you aware of any formal assessments or evaluations of EHR data quality for syndromic surveillance?
5. *Describe findings of literature review briefly and solicit feedback*.
6. Are you aware of any other resources we should be reviewing?
7. Are there other experts you recommend that we speak with?
8. What recommendations do you have for improving these issues or addressing problems?

Appendix C – EHR for Syndromic Surveillance Workgroup Discussion Agendas and Participation

The EHR for Syndromic Surveillance Workgroup met on the following dates and covered the topics listed in Table 4 below. For each meeting, the number of participants attending is provided. The meetings were structured with an overview of the project and then a discussion guided by questions prepared and distributed to members in advance. Meetings were recorded and recordings were available afterward for points of clarification and review of notes.

Table 4. EHR for Syndromic Surveillance Workgroup meeting dates, agendas, and number of participants.

Meeting Date	Agenda	Number of Participants (N=27)
February 26, 2020	Introduction of the team from Thought Bridge Project Overview Progress on the Workplan to Date Literature Review Methods & Findings Discussion including Feedback and Next Steps	24
March 25, 2020	Brief Project Overview Progress on the Workplan to Date Findings of Knowledge Repository Review Questions for Workgroup: 1. What data elements do you consider to be most important for syndromic surveillance? 2. What do you think are the most important aspects of data quality for EHRs used in syndromic surveillance? 3. Are you aware of any formal assessments or evaluations of EHR data quality for syndromic surveillance? 4. Are you aware of any other resources we should be reviewing? Discussion including Feedback and Next Steps	25
April 22, 2020	Brief Project Overview Progress on the Workplan to Date Key Informant Interview Findings Questions for Workgroup: 1. What recommendations/solutions do you have for improving these issues or addressing problems? 2. Who could lead or be involved in implementing these solutions? Discussion including Feedback and Next Steps	27

May 27, 2020	Brief Project Overview Progress on the Workplan to Date Review of Identified Issues 1. What recommendations/solutions do you have for improving these issues or addressing problems? 2. Who could lead or be involved in implementing these solutions? Discussion including Feedback and Next Steps	15
June 24, 2020	Brief Project Overview Vetting Solutions/Participants/Timeline Discussion including Feedback and Next Steps	17
July 22, 2020	Presentation and Discussion of Final Project Findings	14

Appendix D – Table of Proposed Solutions, Leads, and Timeframes.

Table of Proposed solutions, leads, and timeframe to address identified data quality issues.

Solution Area/Solutions	Leads	Timeframe for Implementation
Technical Guidance		
Clarify current PHIN Implementation Guide on CDC NSSP website	CDC	2 months
Archive outdated guides on CDC NSSP website	CDC	2 months
Update/Revise CDC NSSP website text to indicate current Guide (v.2.0) is certifying to 2015 [2]	NSSP	2 months
Determine the feasibility of conducting a review of the current flexibility built into the PHIN Implementation Guide with the aim of achieving a balance between flexibility and data quality	Workgroup or NSSP CoP Data Quality Subcommittee or a workgroup under the subcommittee	9 months
Conduct review of HL7-balloted Messaging Guide, noting any changes/updates/corrections needed and considering addition of new fields needed to address COVID-19	NSSP with the CoP Data Quality Subcommittee, ONC, HL7	9 months after pilot scheduled to begin in October, 2020
Implement enhanced training for sites on using the PHIN Implementation Guide and the HL7 Messaging Guide in tandem, such as a video orientation	NSSP with the NSSP CoP Data Quality Subcommittee	4 months commencing after updates to the Guides
Develop NIST tool tutorial for site to use with facilities	NSSP with the NSSP CoP Data Quality Subcommittee	6 months commencing after updates to the tool
Data Quality Tools		
Update and address issues with the NIST validation tool	NSSP CoP Data Quality Subcommittee or workgroup under the subcommittee with ONC, NIST	6 months
Add the Data Quality Dashboard tools to the 'Staging Phase' of onboarding	CDC with NSSP CoP Data Quality Subcommittee	6-8 months

Share the Data Quality Tools on Demand SAS Program with the ESSENCE Community	CDC with NSSP CoP Data Quality Subcommittee	6-8 months
Provide access to all data quality tools from staging to production phases	CDC with NSSP CoP Data Quality Subcommittee	6-8 months
Data Quality Workshop		
Plan a Syndromic Surveillance Data Quality Workshop involving sites, vendors, facilities and CDC Three proposed objectives/tracks for the workshop: 1. Development of change request/enhancements process and guidelines 2. Development of case studies for successful collaborations, including details related to: a. Networking, coalition building, and strengthening relationships with vendors and facilities b. Coordinating with multiple facility partners – hospital Chief Information Officer (CIO), clinical/admit staff, and IT staff 3. Development of recommendations and best practices to understand, work with, and not disrupt facility workflow while collecting required data elements for syndromic surveillance	Workgroup under the NSSP CoP Data Quality Subcommittee, with CDC and ONC (co-sponsoring), CSTE (hosting)	9-12 months