Data Quality

CONTACT INFORMATION

Submitter name:	Atar Baer, Marcus Rennick
Jurisdiction or affiliation:	Seattle &King County (WA); Marion County (IN)
Phone:	206-263-8154 (Seattle & King County) or 317-221-3362 (Marion County)
Email:	Atar.Baer@kingcounty.gov, mrennick@hhcorp.org

PROBLEM DESCRIPTION

Summarize the problem:
Data collection across a growing stream of contributing facilities and variables requires automated, consistent, and efficient monitoring of quality. Epidemiologists tasked with analyzing syndromic data need to be confident in the overall quality of their data, and aware of the effects of poor data quality when interpreting data. Data quality is also increasingly important as data are shared across jurisdictions and combined for analysis.

The following are significant data quality issues identified by several ISDS- participating jurisdictions:

1. Data drop-offs (variables, sending facilities, etc.)
2. Data completeness/missing data and null values
3. Incorrect values (logical errors, out of range errors, typos, data entry issues, improperly coded values, misclassification)
4. Duplication
5. Hospital/system changes/issues (changes in variables or variable types, changes in coding behavior, hospital additions, field truncation, uninformative generic values, diagnosis in CC field, etc.)
6. Changes in distribution of data

Visualizations and statistical analyses are needed to help syndromic surveillance practitioners identify and summarize the above types of data quality problems.

SOLUTION REQUIREMENTS

Describe the type of solution you are seeking (e.g., anomaly detection, signal validation, data quality characterization):
The solution should be a systematic approach to detecting and visualizing data quality aberrations, addressing as many of the 6 data quality metrics (described above) as possible

The solution should include one or both of the following approaches:

(1) Alerting algorithms to detect changes in data quality. Examples include, but are not limited to, algorithms to identify and characterize significant changes in frequencies of missing or incorrect data values, or algorithms to identify subtle changes in chief complaint coding or diagnosis that may reflect a change in practice at a sending facility. If possible, a standardized weighted data quality score should be developed. This data quality score should be applicable at both a sending facility and jurisdictional level. Volume of patient visits and the severity of the missing variable to analytic interpretation should be considered in the weighing of the final score. For example, missing chief complaint should have greater effect on the data quality score than missing gender. Ideally, an epidemiologist or other analyst would be able to filter data sources based on this data quality score.

(2) Visualizations that allow the practitioner to quickly identify data quality problems across a range of metrics. Examples of visualizations include, but are not limited to, heat maps that display aberrations in data quality, frequency distribution of values, box plots, etc.

Ideally, the output should be organized and presented as a dashboard, or as a single HTML file with hyperlinks to output, so that the practitioner can systematically navigate through the output and easily identify problems across multiple metrics simultaneously.

The solution should strive to accommodate data streams beyond just in the emergency department data setting, allowing the system to scale-up data quality detection for other sources of (near) real-time health surveillance feeds (ie school-absenteeism, outpatient, OTC sales, etc). Visualizations should distinguish between sending facilities and data types, if applicable. Both recent and historical views may be necessary to identify subtle changes in data quality over time.

Describe what type of solution would enable you to implement it in your practice setting (e.g., Do you need an algorithm? Do you need code? If you need code, does it have to be written in any particular programming language?).
The solution should strive for a generalized" plug-and-play" code that can be distributed to multiple jurisdictions with disparate types of data. The solution should be SAS-based and should provide enough documentation within the code so that practitioners can make modifications, where appropriate, to accommodate their particular surveillance needs.

Describe who will use the solution. For example, how many users will there be and what level of skill do the users have? Are the users all within a single jurisdiction/organization?
The solution should be geared toward epidemiologists/analysts with knowledge of the data sources and the ability to correct problems with the data or understand how the identified errors will affect interpretation.

Note any other constraints:
The solution would be a real-time dashboard that updates with the frequency of the incoming data (at least once daily).

VALIDATION

Does a gold standard exist with which to validate the proposed solutions?

- ☐ Gold standard exists within the provided data set (e.g., an outbreak signal nested within baseline data)
- ☐ Gold standard exists in a separate data set, which can be provided to the workgroup (e.g., laboratory data to validate ED data)
- ☐ Gold standard exists but cannot be furnished
- ✔ Gold standard does not exist

INPUT DATA

List the minimum data elements that can be provided to address the problem:

Unique visit ID, Visit Date, Chief Complaint, Facility ID, Age, Gender, 5-digit ZIP, Race/Ethnicity, Disposition, Diagnosis

How much historical data can be provided?

One year of data from both jurisdictions. Data sets have been masked and altered to reduce the risk of releasing protected health information.

Describe any restrictions for sharing the data:

Brief DUAs for both Marion County and Seattle & King County will need to be signed by all institutions working on the data.

Note any other relevant data characteristics:

There are two data sets from two jurisdictions. Methods should be easily applied to both jurisdictions' data.

OUTPUT DATA

NOTES