Data Validation of Health Data in Environmental Health Surveillance

Piloted solutions and lessons learned by the Environmental Public Health Tracking Program

Mackenzie Malone, MPH; Heather Strosnider, PhD, MPH; Mikyong Shin, DrPH, MPH, RN
Environmental Health Tracking Section

NAHDO Annual Conference
August 18, 2020
Outline

- The Environmental Public Health Tracking Program
- Overview of Tracking Data Calls
  - Hospitalizations and Emergency Department Visits Data
- Tracking Validation Process
- What is “Meaningful Difference”? 
- Piloted Solutions
- Summary and Lessons Learned
The Environmental Public Health Tracking Program
Better information for better health

The National Environmental Public Health Tracking Network (Tracking Network) brings together health data and environment data from national, state, and city sources and provides supporting information to make the data easier to understand. The Tracking Network has data and information on environments and hazards, health effects, and population health.

On the Tracking Network, you can:

- Use the Data Explorer to view interactive maps, tables, and charts
- View Info by Location for county level data snapshots
- Visit state & local tracking websites

CDC’s National Environmental Public Health Tracking Program created and maintains the Tracking Network. Learn more about Tracking.
Overview of Tracking Data Calls

- The Tracking Program receives data from recipient states through annual data calls
  - Data is nationally consistent
  - Data dictionaries and How-to Guides
- Data are submitted using a standardized XML schema through Tracking’s secure data submission gateway
- Data thoroughly reviewed by CDC data management unit
Hospitalization and Emergency Department Visits Data

- **Hospitalization (Inpatient Discharge) data:**
  - Asthma
  - Chronic Obstructive Pulmonary Disease (COPD)
  - Carbon Monoxide Poisoning
  - Heat Stress Illness
  - Acute Myocardial Infarction

- **Emergency Department Visits Data:**
  - Asthma
  - COPD
  - Carbon Monoxide Poisoning
  - Heat Stress Illness
High Level Overview of Validation Process

Tier 1
- Grantees submit XML file of data
- Data go through a schema check at the gateway before being received at CDC

Tier 2
- Data that passes the gateway checks are loaded into the database
- Data undergo a series of thorough validation checks

Tier 3
- Valid data are moved to T3 and combined with previously submitted data
- Counts are validated

Tier 4
- The complete dataset is used to calculate the data measures
- Measures are validated and published on the Portal
Tracking Data Validation

- Strange Patterns
- Lack or Excess of Data
- Outliers or Inconsistencies
- Unexpected Results
Unexpected Results – The Archive Comparison Check

- When data are determined to be “too different” from the previous data clarification is requested or the submission fails

- Previous Solution:
  - Count and percent difference thresholds for archive data checks
  - Arbitrary thresholds
  - Most commonly flagged check
  - On average, clarification was needed for over 50% of the submitted files every year

- How do we determine when change in data is due to chance alone or is a true error?
  - The “Meaningful Difference” issue
The Meaningful Difference Problem

- The “meaningful difference” problem:
  - Surveillance data is expected to vary year to year
  - How do we explain what is just expected variation in our hospitalization and ED data and what is error?

- Why this is important:
  - To improve data quality
  - To have confidence in the observed trends
  - To know when public health interventions are needed
Piloted Solutions

Spring 2015: Visual Boxplots

Fall 2016: Tolerance Intervals

Fall 2017: Poisson crude rate comparison

Fall 2018: Standard Deviation Check

Present
Boxplot Visual Trend Check

Spring 2015: Visual Boxplots

Fall 2016: Tolerance Intervals

Fall 2017: Poisson crude rate comparison

Fall 2018: Standard Deviation Check

Present
Box Plot - Results

- **Pros:**
  - Uses all years of data
  - Shows trend
  - Easy to spot outliers
  - Compares summary statistics

- **Cons:**
  - Review of boxplots is manual
  - Results are inferred
  - Not useful for ALL Tracking datasets

- **Has been used for all data calls since implementation and has been adapted for all recipient submitted datasets**
Tolerance Interval Check

- Show the expected range of individual observations
- Allows you to set the confidence (alpha) and percent of population (gamma)
- Set different alpha and gamma values to determine the appropriate threshold
Tolerance Interval - Results

- **Pros:**
  - More statistically sound approach

- **Cons:**
  - Relied on determining arbitrary thresholds
  - Concern of missing records or flagging too many
  - Statistical assumptions
  - Not useful for all Tracking datasets
  - Most reports produced a large output

**Check did not reduce the number of follow ups Tracking was performing throughout the data call**
Poisson Rate Comparison

Spring 2015: Visual Boxplots

Fall 2016: Tolerance Intervals

Fall 2017: Poisson crude rate comparison

Fall 2018: Standard Deviation Check

Present
Rate Comparison - Results

- **Pros:**
  - Uses rates
  - Population denominator helps standardize small counts
  - More rooted in statistics

- **Cons:**
  - Number of counties/records can affect power

- This check in combination with the box plots has been very helpful
- Still being used for validation and has been adapted for all applicable datasets
Standard Deviation Check

- This check uses all previously submitted years of data for a single state and health outcome
- Compares summary statistics from previously submitted data to new years of submitted data
Standard Deviation Check - Results

- Pros:
  - The calculated threshold is dynamic
  - Use of all previous years of data for comparison
  - Focuses on distribution of counts at state and county level

- Cons:
  - Inconsistent with catching errors
  - Less successful with data with small counts (CO Poisoning and Heat Stress Illness)

- This check has been useful to supplement other archive checks
- Provides additional useful information about the distribution of the data
- Helps identify possibly problematic counties
Summary-Improvements in Data Call

<table>
<thead>
<tr>
<th>Metric</th>
<th>Fall 2015</th>
<th>Fall 2019</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Files Received</td>
<td>533</td>
<td>537</td>
</tr>
<tr>
<td>Percent of Submissions requiring follow up</td>
<td>71%</td>
<td>36%</td>
</tr>
<tr>
<td>Time to Public portal</td>
<td>~6 months</td>
<td>~4 months</td>
</tr>
</tbody>
</table>
Summary-Validation Success Story

Before resubmission

After resubmission
Lessons Learned

▪ Hospitalization and emergency department visits data for surveillance poses unique challenges in spotting errors

▪ Exploring and piloting of more sophisticated checks have had mixed results
  • Visual checks have shown effective in spotting errors

▪ The introduction of advanced validation checks have shown to conserve program time and resources

▪ Tracking will continue to review and improve the validation process and pilot solutions to improve accuracy and timeliness of the hospitalization and emergency department visits data
Thank you!