Preparation of Round 1 LT2 Monitoring Data for EPA Analyses

Introduction

This document describes the quality assurance (QA) and quality control (QC) steps that were taken on the first round of Long-Term 2 Enhanced Surface Water Treatment Rule (LT2) Cryptosporidium monitoring data as the data were uploaded into the Data Collection and Tracking System (DCTS). The document also describes data that may have data quality concerns based on comments provided with the data. In addition, the document includes a list of data that EPA may exclude from any future analyses of the data. EPA included these descriptions so users of these data are aware of potential data quality concerns. EPA developed a 'cleaned up dataset', which provides the data remaining after all data quality concerns, unnecessary data fields, and added data fields described below have been addressed. However, the data of potential concern remain in the publicly available original database so that data users can retain the data for their analyses if so desired.

The original database contains more than 40,000 records for *Cryptosporidium* and more than 50,000 records for *Escherichia coli*. Not all LT2 monitoring data are in the original database. Not included are some grandfathered data and data from systems that did not report the data. (Systems serving fewer than 10K were not required to submit data to DCTS, though some did.)

Data from the DCTS database also contains a list of 252 facilities that indicated they would opt for maximum treatment in lieu of *Cryptosporidium* monitoring. Additional facilities may have also opted to provide maximum treatment without entering that information in DCTS.

DCTS *Cryptosporidium* and *E. coli* Data

Cryptosporidium: http://water.epa.gov/lawsregs/rulesregs/sdwa/lt2/upload/cryptodatareported.csv
E. coli: http://water.epa.gov/lawsregs/rulesregs/sdwa/lt2/upload/ecolidatareported.csv

QA/QC Steps

Data pulled from DCTS were to have undergone a series of QC checks and QA oversight by labs and water systems.

- Labs attest to the quality of data as they enter it into DCTS.
- Systems are notified as each of their records is approved (by the lab) and are expected to review and comment.
- After review and approval (or after a fixed period of time in with no response from the system), the data are elevated so regulators (e.g., EPA and states) may see the data. Prior to this step, regulators do not have access to the data.

Data Entries with Potential Data Quality Concerns

The data include thousands of records with comments. Most of these comments are innocuous, but some may raise questions about the quality of the data. Information in other fields that may have bearing on data quality include:

- For *Cryptosporidium*,
  - 107 records with VALID_STATUS_CODE "returned"
  - 62 records with VALID_STATUS_CODE "contested,"
  - 2 records with VALID_STATUS_CODE "epa contested"
  - 14,364 records with VALID_STATUS_CODE "not reviewed."
  - LT2DB_STG2_FACILITY_STATUS is "Inactive" for 55 records.
  - The single most questionable record (a sample for with an oocyst count of 216,680) received no comments and has VALID_STATUS_CODE "not reviewed."
- For *E. coli*,
  - 128 records with VALID_STATUS_CODE "returned"
  - 77 records with VALID_STATUS_CODE "contested,"
  - 4 records with VALID_STATUS_CODE "epa contested"
  - 18,807 records with VALID_STATUS_CODE "not reviewed."
  - LT2DB_STG2_FACILITY_STATUS is "Inactive" for 42 records.

Data Potentially Excluded from Future EPA Analyses

The following *Cryptosporidium* records (rows) are in the original dataset, but EPA may exclude them from our analyses for the following reasons:

For *Cryptosporidium*,

- 1 record with SAMPLE_CRYPTO_ID 40670. Per follow-up communication by EPA, the *Cryptosporidium* count of 216,880 was an entry error. The correct entry is "0".
- 2 records with SAMPLE_CRYPTO_IDs 18346 and 18347. Per EPA_FLAG_EXPLANATION, "EPA agrees that the 2/27/07 crypto results are not valid and should be removed."
- All 297 records associated with PA1460073 may be excluded from EPA's analyses because the counts and volumes used to represent flow weighted average concentrations were incorrect as entered.
- All 21 records associated with OH5501211 may be excluded from EPA's analyses because the counts and volumes used to represent flow weighted average concentrations were incorrect as entered.
- Any matrix spike having less than 80 Cryptosporidium spiked may not be included in EPA's recovery modeling.
- *Cryptosporidium* matrix spike records having significantly more oocysts counted than spiked.
- The two samples with VALID_STATUS_CODE "epa contested" will not be included in EPA's analyses.
- *Cryptosporidium* and *E. coli* records that were duplicates of other records (redundant)

The following data fields (columns) are in the original dataset, but EPA may exclude them from our analyses for the following reasons:

- The DATA_QUALIFIER_FLAG field was removed because it was blank for all *Cryptosporidium* records.
- The ORGANIZATION_NAME field was dropped because ORGANIZATION_CODE uniquely identifies the organization (public water system).
- The LT2DB_STG2_PWS_DETAILS_STATUS field because it is listed as "Active" for all *Cryptosporidium* records.
- Information from the following fields (LAB_APPROVAL_DATE, LAB_COMMENT, PWS_REVIEW_DATE, etc.) have been reviewed and information relevant for the analyses have been moved to new "Comment" fields. The original fields were removed.

The following information has been included in the cleaned up dataset to clarify some potential data quality issues:

- The *E. coli* field EcoliQual and EcoliNumeric report the sign and numeric value for *E. coli* concentrations reported with "less than" (<) or "greater than" (>) values. For example "> 100" was separated into a qualifier of EcoliQual = ">" and a value of EcoliNumeric = 100.
- *E. coli* and *Cryptosporidium* "Comment" fields include notations regarding data quality.
- The *Cryptosporidium* field "Actual_Count" indicates whether the number of oocysts counted (NO_OF_CRYPTO) is an actual count or a value derived to represent a weighted average of concentrations measured in two different source waters.
- The *E. coli* field "Conc_per_100_ml" combines concentrations reported in field SAMPLE_CALC and concentrations derived from raw data.