NEON ALGORITHM THEORETICAL BASIS DOCUMENT: TOS MOSQUITO ABUNDANCE AND DIVERSITY - QA/QC OF RAW FIELD AND LAB DATA

PREPARED BY	ORGANIZATION	DATE
Katherine LeVan	FSU	05/03/2016
Kimberly Tsao	FSU	05/03/2016

APPROVALS	ORGANIZATION	APPROVAL DATE

RELEASED BY	ORGANIZATION	RELEASE DATE
Jen DeNicholas	CM	mm/dd/2015

See configuration management system for approval history.

The National Ecological Observatory Network is a project solely funded by the National Science Foundation and managed under cooperative agreement by NEON, Inc. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

CHANGE RECORD

REVISION	DATE	ECO#	DESCRIPTION OF CHANGE
A	mm/dd/yyyy	ECO-xxxx	Initial Release
B	mm/dd/yyyy	ECO-xxxx	
C	mm/dd/yyyy	ECO-xxxx	

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

1 DESCRIPTION

1.1 Purpose

This document details the algorithms used for creating a subset of NEON Level 1 data products that are the quality controlled products generated from raw Level 0 data, and associated metadata. In the NEON data products framework, the raw data collected in the field, for example, a negative test result for a particular pool of mosquitoes, are considered the lowest level (Level 0). Raw data that have been quality checked via the algorithms detailed herein, as well as simple metrics that emerge from the raw data, such as total species richness of mosquitoes at a particular site, are considered Level 1 data products. This document relates only to the former group of L1 data products, the quality controlled pass-through products from the Level 0 data products.

It includes a detailed discussion of measurement theory and implementation, appropriate theoretical background, data product provenance, quality assurance and control methods used, approximations and/or assumptions made, and a detailed exposition of uncertainty resulting in a cumulative reported uncertainty for this product.

1.2 Scope

This document describes the theoretical background and entire algorithmic process for creating a subset of quality controlled and calibrated L1 data products and associated metadata from input data. These data products include mosquito trapping data (NEON.DOM.SITE.DP1.10043.001) and pathogen status (NEON.DOM.SITE.DP1.10041.001). It does not provide computational implementation details, except for cases where these stem directly from algorithmic choices explained here. This document also provides details relevant to the publication of the data products via the NEON data portal, with additional detail available in the NEON Data Publication Workbook for NEON Data Publication Workbook for TOS Mosquito Abundance, Diversity and Phenology (AD[09]).

This document describes the algorithms for ingesting and performing automated quality assurance and control procedures on the data collected in the field pertaining to Field and Lab Protocol: Mosquito Sampling (AD[11]). The raw data that are processed in this document are detailed in the NEON Raw Data Ingest Workbook for TOS Mosquito Abundance and Diversity (AD[08]).

1.3 Acknowledgements

Dr. David Hoekman and Dr. Yuri Springer contributed to an early version of this ATBD.

2 RELATED DOCUMENTS AND ACRONYMS

2.1 Applicable Documents

Applicable documents contain information that shall be applied in the current document. Examples are higher level requirements documents, standards, rules and regulations.

AD[01]	NEON.DOC.000001	NEON Observatory Design (NOD) Requirements
AD[02]	NEON.DOC.002652	NEON Level 1, Level 2, and Level 3 Data Products Catalog
AD[03]	NEON.DOC.005005	NEON Level 0 Data Product Catalog
AD[04]	NEON.DOC.005011	NEON Coordinate Systems Specification
AD[05]	NEON.DOC.004309	NEON Field Site Information
AD[06]	NEON.DOC.002261	TOS Spatial Data
AD[07]	NEON.DOC.tax	Master Taxonomy for Mosquitoes
AD[08]	NEON.DOC.001401	NEON Raw Data Ingest Workbook for TOS Mosquito Abundance, Diversity and Phenology
AD[09]	NEON.DOC.001412	NEON Data Publication Workbook for TOS Mosquito Abundance, Diversity and Phenology
AD[10]	NEON.DOC.000910	TOS Science Design for Mosquito Abundance, Diversity and Phenology
AD[11]	NEON.DOC.000911	TOS Science Design for Vectors and Pathogens
AD[12]	NEON.DOC.014049	TOS Protocol and Procedure: Mosquito Sampling

2.2 Reference Documents

Reference documents contain information complementing, explaining, detailing, or otherwise supporting the information included in the current document.

RD[01]	NEON.DOC.000008	NEON Acronym List
RD[02]	NEON.DOC.000243	NEON Glossary of Terms

2.3 External References

N/A

2.4 Acronyms

N/A

3 DATA PRODUCT DESCRIPTION

Mosquito sampling shall be conducted at regular intervals by NEON field technicians at core and relocatable sites. When adult mosquitoes are active, sampling bouts will occur every two weeks at core sites and every four weeks at relocatable sites. For additional details on the sampling design and associated protocol, see the NEON Science Design for Mosquito Abundance, Diversity, and Phenology (AD[10]), the NEON Science Design for Vectors and Pathogens (AD[11]) and the TOS Protocol and Procedure: Mosquito Sampling(AD[12]).

Mosquito-borne pathogen sampling involves the testing of all or a subset of collected mosquitoes for infection by viral pathogens by one or more external facilities. Test results and associated documentation will be provided to NEON and used to report the prevalence of pathogens. Testing will yield data on the presence of important mosquito pathogens (e.g., West Nile virus, Eastern equine encephalitis virus, dengue, etc) in a subset of species that are known vectors of disease. See the TOS Science Design for Vectors and Pathogens (AD[11]) for additional background on mosquito-borne pathogen sampling.

3.1 Variables Reported

This ATBD describes the steps needed to generate the mosquito-related L1 data products, mosquito sampling data (NEON.DOM.SITE.DP1.10043) and mosquito pathogen-status (NEON.DOM.SITE.DP1.10041). Mosquito sampling data provide information as to the species identity and estimated abundance of mosquitoes sampled from CO_2 light traps (see Figure 1 for details). Mosquito-borne pathogen status data relate to test results for the presence of pathogens in each tested mosquito pool (see Figure 1 for details).

Subproducts for this data product are listed below. Detailed lists of the associated subproducts and metadata products are provided separately, along with example data in publication-ready spreadsheets in the NEON Data Publication Workbook for TOS Mosquito Abundance, Diversity and Phenology (AD[09]). Field names have been standardized with Darwin Core terms (http://rs.tdwg.org/dwc/; accessed 16 February 2014), the Global Biodiversity Information Facility vocabularies (http://rs.gbif.org/vocabulary/gbif/; accessed 16 February 2014), the VegCore data dictionary (https://projects.nceas.ucsb.edu/nceas/projects/bien/wiki/VegCore; accessed 16 February 2014), and with the Bird Monitoring Data Exchange standards (http://www.avianknowledge.net; accessed 16 February 2014), where applicable. Geospatial data shall conform to the standards set forth in the NEON Coordinate Systems Specification (AD[05]).

The data products result from the field collection of CDC CO_2 light traps and lab processing of samples. Mosquitoes will be taxonomically identified by external facilities (to species whenever possible) and identifications for a subset of difficult taxa will be verified by DNA barcoding.

Number	Field Name	Description
TBD	scientificName	Scientific name, associated with the taxonID. This is the name of the lowest level taxonomic rank that can be determined
TBD	estimatedAbundance	Estimated number of individuals per trap
TBD	testResult	Result of the pathogen test

3.2 Temporal Resolution and Extent

The finest temporal resolution at which mosquito data (for the purposes of species richness, abundance and phenology) will be tracked is trapping night or trapping day. The finest level of temporal resolution at which mosquito-borne pathogen status will be tracked is at the level of a sampling bout. Two trapping nights and the intervening day for up to ten plots comprise a sampling bout of three separate samples per trap (30 samples per site; see Figure 1). The setDateTime (indicating when the trap was set) and collectDateTime (indicating when the trap was collected)

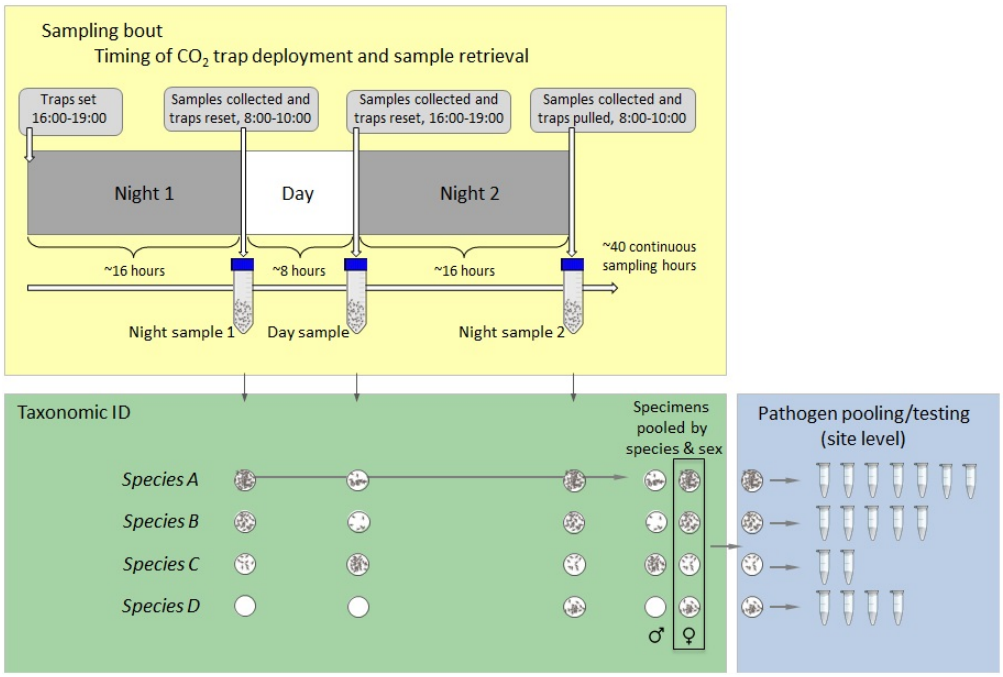

Figure 1: A workflow illustrating the process of data collection for the mosquito protocol

will be recorded for each sample collected during a bout. Bouts are grouped using the **eventID** designation (a descriptor that includes the year of sampling, the site ID, and the calendar week in which a sampling bout occurred).

The total number of bouts per year varies among sites based on seasonality of each site (e.g., stopping during winter at temperate sites). During the time of year when mosquitoes are flying, sampling bouts occur every 2 weeks at the core site and every 4 weeks at each relocatable site, alternating between the core and a relocatable such that one site in the domain is sampled each week. After the mosquito season has ended (e.g., upon the onset of winter), weekly sampling at the core site will monitor for mosquito presence and help determine when the next mosquito sampling season should begin. A given sampling bout will be cancelled if minimum ambient temperature thresholds are not met. Additional details about sampling bout frequency can be found in the TOS Protocol and Procedure: Mosquito Sampling (AD[12]).

Finally, while the temporal unit of bout will be used for calculating abundance estimates and other higher level data products, data may also be aggregated to alternative temporal resolutions (e.g., 2-night bout, month, season) by end users.

3.3 Spatial Resolution and Extent

The finest spatial resolution at which mosquito sampling data will be tracked is plot (one trap per plot sampled 3 times per bout; see Figure 1) for diversity, abundance and phenology metrics. The finest level of spatial resolution at which mosquito-borne pathogen status will be tracked is at the level of the sampling site because pathogen testing data are pooled across the entire site for pathogen testing. Additional details about sample pooling and testing can be found in the NEON Science Design for Vectors and Pathogens (AD[11]). Plots are placed in order to

be representative of a site (e.g., spread across dominant vegetation types). Three sites, a core and two relocatables, are together representative of a domain. Every mosquito specimen can be traced to a specific collection event (collectDateTime) and plot location (plotID) where it was collected. Overall, this results in a spatial hierarchy of:

plotID (finest spatial resolution) < siteID (spatial aggregation of multiple plotIDs) < domainIDs (spatial aggregation of multiple siteIDs)

3.4 Associated Data Streams

Mosquito-borne pathogen status data are directly linked to data from mosquitoes sampled from CO_2 traps AD[09]. This linkage will occur through the variables **eventID** and **scientificName**. These variables combined with **individualCount** in the mos_identification_pub indicate the number of female mosquitoes of a particular species available for pathogen testing. Summing **poolSize** for records with the same **eventID**, **scientificName** in mos_pathogenresults_pub will indicate the number of mosquitoes that were tested for pathogens. A third table, mos_archival_pub, will provide information as to how many of the non-tested mosquitoes from a given bout (records with the same **eventID** and **scientificName**) were archived.

There will be linkages between the data products described in this document and others that are produced by NEON. For instance, NEON is collecting data on bird abundance and diversity in areas being sampled for mosquitoes. The co-location of sampling allows end users to make interesting connections between taxa (e.g., between the presence of certain bird taxa and the appearance of West Nile within mosquitoes).

3.5 Product Instances

Mosquito traps will be set at all core and relocatable sites. Up to 10 traps will be set at every site. Mosquito density varies dramatically, and trapping could result in as few as zero or as many as tens of thousands of individuals per trap. A small number of mosquitoes will be mounted on paper tabs on pins (called pointing). The number of pointed mosquitoes will be limited to only those that are submitted for DNA barcoding. It is expected that 50-150 mosquitoes per domain per year will be selected for barcoding.

4 Scientific Context

4.1 Theory of Measurement/Observation

4.2 Mosquitoes as sentinel taxa

The Terrestrial Observation System (TOS) is charged with monitoring the responses of biodiversity and ecosystems to environmental change. While several different invertebrate groups were considered, a NEON design committee (AIBSnews, 2007) selected mosquitoes (Diptera: Culicidae) as a focal taxon for measurement. Mosquitoes are a diverse and widespread family of insects that have been extensively studied because of their ecological and epidemiological significance. From an ecological standpoint, mosquitoes are a dominant taxon in aquatic food webs, due to aquatic larval and pupal forms and the reliance of adults on water for breeding habitat. As such, mosquitoes

compose a sizable proportion of invertebrate biomass in aquatic systems and act as a key food source for aquatic and terrestrial predators (e.g., fish, amphibians, spiders, birds). As a result, decreases in mosquito populations, e.g., in response to land-use alteration or climate change, may have widespread negative impacts on ecosystems. From an epidemiological standpoint, mosquitoes act as vectors for numerous parasites and pathogens that may have marked impacts on humans, livestock [e.g., Rift Valley Fever (Daubney and Hudson 1931), Eastern equine encephalitis (Kissling et al. 1954)], and wildlife [e.g., avian malaria (van Riper et al. 1986), West Nile virus (Marra et al. 2004)]. For example, the emergence of West Nile virus in North America has resulted in widespread population declines of several common birds (e.g., crows, robins, wrens, chickadees, blue jays; LaDeau et al. 2007) with important potential consequences for ecosystem services like seed dispersal, carrion scavenging and insect regulation (LaDeau et al. 2008). It is due to their implications for human health, however, that mosquito biology and ecology have been most extensively studied, in order to characterize how mosquitoes spread pathogens and mitigate the impacts of mosquito-borne diseases. Furthermore, and for the same reasons, populations have been and continue to be monitored by national, state and local agencies.

The distribution and seasonal phenology of mosquito populations are influenced by many landscape factors, including climate, vegetation and host availability (Buckner et al. 2010, Reisen 2010). As a result, mosquitoes are highly sensitive to environmental gradients and perturbations. In addition, the short generation time and high fecundity of these insects allow them to respond quickly to environmental change. Together, these factors make mosquitoes an ideal sentinel taxon for evaluating the ecological effects of global change phenomena.

4.3 Mosquitoes as a vector of disease

Worldwide, mosquito-borne pathogens are responsible for a human health burden unsurpassed among vector-borne diseases. In 2004 alone over 1.8 million human deaths were attributed to malaria (Murray et al. 2012), and 96 million people are estimated to experience disease associated with infection by dengue viruses each year (Bhatt et al. 2013). Moreover, mosquito-borne pathogens can also cause substantial reductions in populations of livestock and wildlife, with potentially important repercussions for human health, economic productivity, and the structure and function of ecological communities (e.g., Daubney et al. 1931, van Riper et al. 1986, Morris 1989, Scott and Weaver 1989, LaDeau et al. 2007, Paweska and van Vuren 2014).

Forecasts of potential ecological and public health consequences of climate change often focus on mosquitoes and the pathogens they transmit (Shope 1992, Reeves et al. 1994, Sutherst 2004, Harrigan et al. 2014). Although mosquitoes occur worldwide, they are most consistently abundant in localities with tropical or moderately temperate climes in which relatively warm and wet conditions prevail (although populations of some species in subarctic and alpine regions reach extremely high abundance during parts of the year). As a result, increases in temperature or precipitation at higher latitudes or elevations due to climate change could promote range expansions of mosquitoes currently confined to tropical areas (Epstein et al. 1998, Patz et al. 2000, Caminade et al. 2012, Eisen and Moore 2013, but see Reiter, 2001). This geographic spread could be facilitated by the periodic long-distance transport of mosquitoes that can occur incidentally as part of human travel and international commerce (Lounibos 2002, Tatem et al. 2006). Establishment and spread of mosquito species into new localities creates the potential for associated pathogens to be concurrently introduced (e.g., Gould and Higgs 2009, Weaver and Reisen 2010). Additionally, there is abundant evidence that changing climatic conditions will significantly affect mosquito demography and processes associated with the transmission of mosquito-borne pathogens (Mordecai et al. 2013). For example, changes in ambient temperature are predicted to alter mosquito vectorial capacity (Watts et al. 1987, Reisen et al. 2006, Paaijmans et al. 2012) and biting rates (Lardeux et al. 2008), and may in some cases catalyze host range

shifts in arboviruses (Brault and Reisen 2013). Because of their extensive geographic distribution, ecological and epidemiological significance, and sensitivity to processes associated with climate and land-use change, mosquitoes and the pathogens they transmit are natural targets for NEON sampling.

Collected mosquitoes will be sent to one or more external facilities for taxonomic identification and pathogen testing. Mosquitoes will be enumerated by species and sex for each sampling bout at each plot at each site (e.g. number of female mosquitoes of species A collected at plot B within site C during sampling bout D). Following identification, mosquitoes will be combined by species, sex, and bout (e.g., all female mosquitoes of species A collected at site C during sampling bout D), and a subset will be tested for infection by viral pathogens. Multiple individuals will be pooled for testing, which may include RT-PCR, Vero cell culture, and melt curve assays. These methods vary in target specificity, from general (e.g., Vero cell culture) to specific viral species (e.g., RT-PCR).

4.4 Theory of Algorithm

This document describes the algorithms for assessing the integrity of the L0 data stream generated by the field sampling and lab analysis of mosquito abundance and diversity. The approaches described herein are simple yet necessary components of quality control and quality assurance, including: verifying that all required data are recorded for each trapping event, comparing data values to pre-defined (through provided validation rules or specified lookup tables) ranges, and tracking the individual-level data for consistency and accuracy through time. A schematic overview of relationships among L0 and L1 products is provided in Figure 2.

4.5 Special Considerations

The mosquito data are unusual among the TOS data products in that many specimens will be destroyed during the pathogen testing that follows identification.

5 Data Entry Constraint and Validation

Many quality control measures may be implemented at the point of data entry. Constraining data formats (in spreadsheets and MDR or webUI fields) and providing dropdown options on MDR and/or webUI applications reduces the number of processing steps necessary to check the raw data and prepare it for publication. This section describes the data constraint validation requirements that should be built in to the programs field technicians will uses to record data.

5.1 Run the following steps for all data ingested via the MDR

1. Constrain entered values to the correct **dataType**
2. Constrain entered values to conditions specified by **entryValidationRules**
3. Generate a unique ID (**uid**) for each record
4. Follow guidelines for fields in which no data have been entered, as specified in **noDataOutcomePDA** and **noDataOutcomeUI**:

    a. IF **noDataOutcome** == fail:

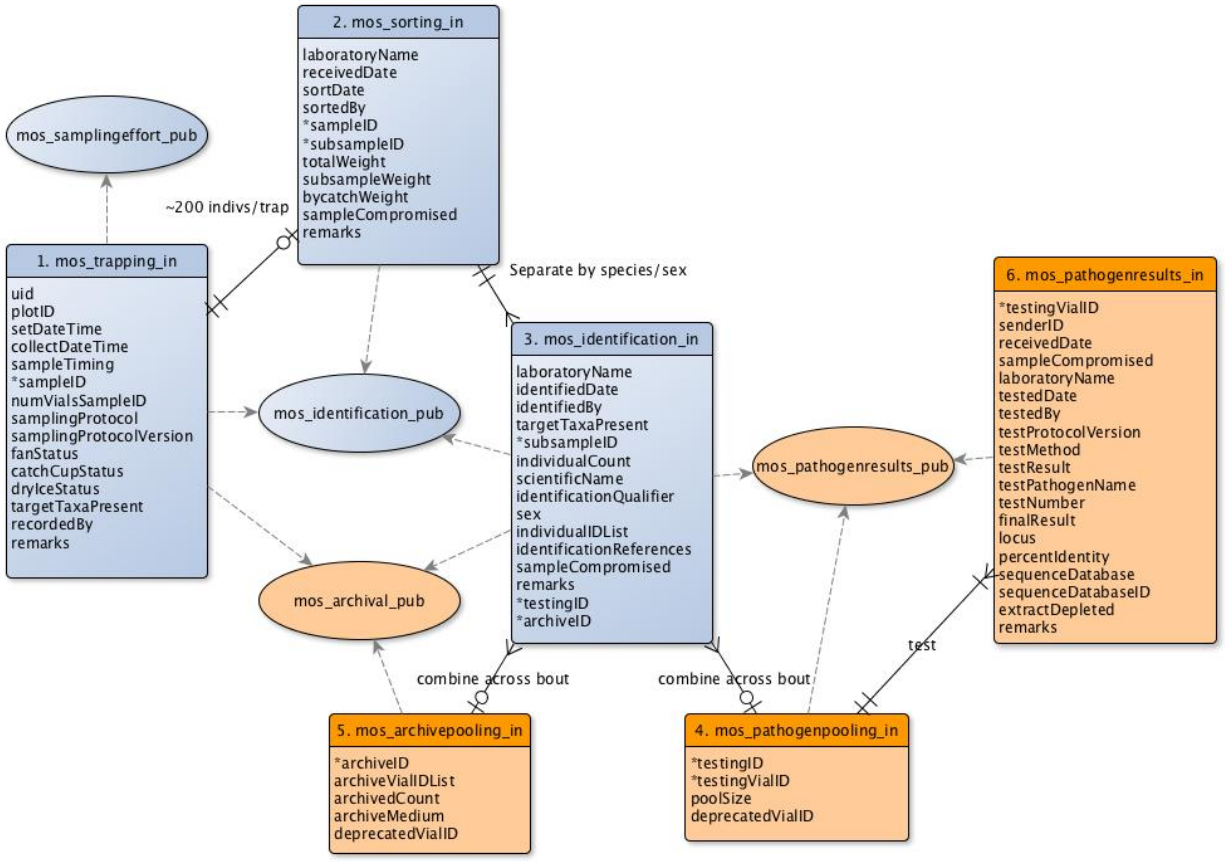

Figure 2: A diagram illustrating the relationships among ingest tables (rectangles) and pubs (ovals, field names not shown). Fields joining tables are indicated with '*'. Join types are indicated using crow's foot notation. Colors indicate spatiotemporal scale: blue indicates trap-bout level, orange indicates site-bout level.

i. do not let user finalize record until a value is provided
ii. warning message text, 'Please enter a value for [fieldname] to continue', unless an alternative is provided in **warningText**

b. ELSE IF **noDataOutcome** == warn:
   i. warn user that a value is missing prior to finalizing record, but allow selection of 'OK' to continue without a value
   ii. warning message text, 'Please confirm that there is no for [fieldname] to continue', unless an alternative is provided in **warningText**

c. ELSE IF **noDataOutcome** == pass:
   i. allow user to finalize record with no values in this field

5. Follow guidelines for **case** and default values specified in **defaultValuePDA** and **defaultValueUI**

5.2 Run the following steps for mos_trapping_in

1. Provide drop-down menu for **domainID**
    a. For subsequent record, auto-populate **domainID** with value entered previously
2. Provide drop-down menu of **siteID**, filtered by the selected **domainID**
    a. For subsequent record, auto-populate **siteID** with value entered previously
3. Provide drop-down menu of **plotID**, filtered by the selected **siteID**
    a. To generate menu of **plotIDs**, use the TOS plot-level spatial data lookup table (AD[06]) and where 'mos' is an element of **applicableModules**
4. Provide drop-down menu of options for **samplingProtocolVersion**
5. **recordedBy** - Defaults to the user logged in to the Mobile Data Recorder (MDR) app or web user interface when the record is created.
    a. MDR will send values from recordedBy as entered on the MDR to both **recordedBy** and **enteredBy** in MDR
6. **measuredBy** - Maximo user list for FOPs or 'Other'
    a. MDR solution: just type in (i.e., typeahead field) rather than select other
7. Provide filtered drop-down menu and/or typeaheads for controlled lists for all fields except **remarks**.
8. Provide a calendar and time field for technicians to add **setDateTime** and **collectDateTime**
    a. **setDateTime** should always precede **collectDateTime** by at least 3 hours (if that is not true; warn the user)
    b. **setDateTime** and **collectDateTime** should not differ by more than 24 hours for any one entry (if that is not true; warn the user)
9. Provide a drop-down menu of **targetTaxaPresent**
10. IF **targetTaxaPresent** is NOT 'N':
    a. Generate a value for **sampleID**
    b. Create a sample table entry with tagID = **sampleID** ELSE IF **targetTaxaPresent** == 'N': **sampleID** is blank; no sample table entry is created

5.3 Run the following steps for all data ingested via spreadsheet

1. Strip all starting and ending double quotes and convert inner quotes to single quotes
2. Strip all leading and trailing white space
3. Fail data ingest if non-ASCII characters are encountered, except where such characters are specified as allowed, in **entryValidationRules**
4. Fail data ingest if all expected columns and column names are not present (even if the column is entirely blank, as may occur if **noDataOutcome** is 'pass' for the field), except where the ingest workbook and ATBD specify that fields are to be generated by CI on ingest (e.g. **uid**). Ignore extra columns, from those expected, and do not ingest such columns

5. If **noDataOutcome** is 'pass', evaluate rules specified in **entryValidationRules** only for non-empty cells
6. Follow AD[08] for requirements on:
    a. Required fields (**noDataOutcomeUI**)
    b. **datatype**
    c. **entryValidationRules**
    d. LOVs (match case-insensitive, but post to database in case specified in the LOV)
7. If there are errors on upload, return the following information to the user: row number; field name; brief description of what failed. If multiple locations in the file fail, print all messages
8. Unless otherwise specified, follow all other rules in the Spreadsheet Ingest Global Rules

5.4 Sample Creation Rules

1. If parent sample is null or not found, then the child sample must be null.
2. For samples that are mixtures (mos_identification_in:testingID, mos_identification_in:archiveID))
    a. For the first record containing a particular ID, create a new sample
    b. For subsequent records containing that ID, do not create a new sample, but add an additional parent (mos_identification_in:subsampleID) to the existing ID.

5.5 Transition Schedule Rules

Data can be transitioned when the data for a complete bout conducted at a given site are returned from Field Operations and external lab facilities. However, each published data table only requires a subset of all ingest tables.

1. mos_samplingeffort_pub and mos_identification_pub both require:
    a. mos_trapping_in
    b. mos_sorting_in
    c. mos_identification_in
2. mos_archival_pub requires:
    a. mos_trapping_in
    b. mos_sorting_in
    c. mos_identification_in
    d. mos_archivepooling_in
3. mos_pathogenresults_pub requires:
    a. mos_trapping_in
    b. mos_sorting_in
    c. mos_identification_in
    d. mos_pathogenpooling_in
    e. mos_pathogenresults_in

Information about each record in the published data comes from one of 6 separate tables generated both internally and externally. As a result, data may be transitioned to create a given publication either when all required data on sampleIDs for specimens generated by Field Operations are also returned by the external identification and pathogen-testing facilities (as needed, see list above for details) or 18 months after Field Operations returns data on a **sampleID**, whichever comes first.

6 Algorithm Implementation

Throughout the algorithm implementation section of this ATBD, 'nodata', 'null', and/or 'NA' indicates a blank cell. All variables reported from the field or laboratory technician (L0 data) are listed in the data ingest workbook (AD[08]), notated here as mos_tablename_in. Unless otherwise specified in the algorithm below, all variables that appear in tables mos_tablename_pub (L1 data) have been passed directly from the L0 variables with the same name, as listed in the data publication workbook (AD[09]). Algorithm implementation should proceed in the order of the subsections provided here, e.g., the processing of data in subsection 6.1 should occur prior to that of data in subsection 6.2, and so on. Table indicate portions of the data only.

6.1 Summary of Algorithm for trapping Data

1. Tidy formatting in all string fields
2. Remove and/or flag duplicate values
3. Rename fields used across multiple tables
4. Add spatial metadata from the TOS lookup table
5. Add field **eventID** (requires a calculation of week number)
6. Add field **trapCompromisedQF** field using **fanStatus**, **catchCupStatus**, and **dryIceStatus** fields
7. Add field **nightOrDay** based on **setDateTime** and **collectDateTime**
8. Add field **trapHours** based on **setDateTime** and **collectDateTime**

6.2 Summary of Algorithm for sorting Data

1. Tidy formatting in all string fields.
2. Remove and/or flag duplicate values
3. Rename fields used across multiple tables

6.3 Summary of Algorithm for identification Data

1. Tidy formatting in all string fields
2. Remove and/or flag duplicate values
3. Rename fields used across multiple tables
4. Append taxonomic information

6.4 Summary of Algorithm for pathogenpooling and archivepooling Data

1. Remove and/or flag duplicate values

6.5 Summary of Algorithm for pathogenresults Data

1. Tidy formatting in all string fields
2. Remove and/or flag duplicate values
3. Rename fields used across multiple tables

6.6 Steps to run for all ingest tables

1. Remove extraneous characters from the **remarks** field.

a. Trim (remove) all leading/trailing spaces, line endings, tabs, and ctr-R in all string fields
b. Replace all non-leading/trailing line endings, tabs, ctr-R in all string fields with a single space
c. Replace any non-ASCII characters in all string fields with a single space
d. Replace any instances of 2+ adjacent whitespace in all string fields with a single space

2. Handling near-complete duplicates: For multiple records identical in all fields (including all sample information for the record, for both primary and child samples) except **uid** and **remarks**: keep one record, concatenating unique values from the **remarks** field, separated by pipes.

Table 2: Partial input table prior to removal of duplicates

plotID	collectDateTime	setDateTime
OSBS_056	20140409.0920	20140408.1846
OSBS_056	20140409.1725	20140409.0920
OSBS_056	20140410.1103	20140409.1725
OSBS_056	20140409.0920	20140408.1846

plotID	collectDateTime	setDateTime
OSBS_056	20140409.1725	20140409.0920
OSBS_056	20140410.1103	20140409.1725
OSBS_057	20140409.0940	20140408.1900

Quality flags indicating duplicates should be made for each ingest table:

mos_trapping_in:**duplicatesInTrappingQF**,

mos_sorting_in:**duplicatesInSortingQF**,

mos_identification_in:**duplicatesInIdentificationQF**,

mos_pathogenpooling_in:**duplicatesInArchivePoolingQF**,

mos_archivepooling_in:**duplicatesInPathogenPoolingQF**,

mos_pathogenresults_in:**duplicatesInPathogenResultsQF**.

IF the record was not checked (any required field is NA): 'duplicateQF' = -1

IF there are no duplicates of a record: 'duplicateQF' = 0

IF duplicates were found for a record: 'duplicateQF' = 1

Table 3: Partial output table after removal of duplicates

plotID	collectDateTime	setDateTime	duplicatesInTrappingQF
OSBS_056	20140409.0920	20140408.1846	1
OSBS_056	20140409.1725	20140409.0920	0
OSBS_056	20140410.1103	20140409.1725	0
OSBS_056	20140409.1725	20140409.0920	0

Handling partial duplication. Certain combinations of values within a table are deviations from the protocol and likely represent partial duplicates rather than independent records. In some cases, these partial duplicates are resolveable programatically and the resolved record can be marked with a 'duplicateQF' flag = 1. In other cases, no resolution will be possible and the 'duplicateQF' flag will = 2.

3. The mos_trapping_in table should not have more than one record for a unique **collectDateTime** and **plotID** combination. These 2 fields are the key to detect partial duplicate records within mos_trapping_in.
    a. For records that conflict in values for fields in the list of resolveable conflicts (specified below), but are otherwise identical; keep one record and populate the conflicted fields with values as specified below in the list of resolveable conflicts:

List of resolveable conflicts:

setDateTime = the latest **setDateTime** among values entered,

recordedBy = value from the last record entered,

sampleProtocolVersion = the latest version among values entered,

sampleTiming = "Field season",

fanStatus = 'Off',

catchCupStatus = 'Disturbed' if one of the values recorded, otherwise 'Missing',

dryIceStatus = 'Absent',

targetTaxaPresent = 'Y';

For the record kept, **duplicatesInTrappingQF** = 1; concatenate **remarks** as described above.

b. If multiple records still share identical **collectDateTime** AND **plotID** values after following step a (above) and have conflicting values or sample information for any field not specified in part a (i.e., the list of resolveable conflicts), then these records are unresolveable. In that case, retain both records and assign both with **duplicatesInTrappingQF** = 2

4. The mos_identification_in table should not have more than one record per unique combination of **scientificName**, **sex**, **subsampleID**, and **identificationQualifier**. These 4 fields are the key to detect partial duplicate records within mos_identification_in.

    a. For records that conflict in values for fields in the list of resolveable conflicts (specified below), but are otherwise identical; keep one record and populate the conflicted fields with values as specified below in the list of resolveable conflicts:

List of resolveable conflicts:

identifiedDate = the latest **identifiedDate** among values entered,

targetTaxaPresent = 'Y',

individualIDList = a pipe-concatenated list of all elements within **individualIDList** from both records,

sampleCompromised = the **sampleCompromised** value that is not "No known compromise",

identificationReferences = the latest **identificationReferences** among values entered,

identifiedBy = the latest **identifiedBy** among values entered,

For the record kept, **duplicatesInIdentificationQF** = 1; concatenate **remarks** and **identificationRemarks** as described above.

b. If multiple records still share identical **scientificName**, **sex**, **subsampleID**, AND **identificationQualifier** values after following step a (above) and have conflicting values or sample information for any field not specified in part a (i.e., the list of resolveable conflicts), then these records are unresolveable. In that case, retain both records and assign both with **duplicatesInIdentificationQF** = 2

5. The mos_pathogenresults_in table should not have more than one record per unique **testingViallD**, **testNumber**, and **testPathogenName** combination. These 3 fields are the key to detect partial duplicate records within mos_pathogenresults_in.

    a. For records that conflict in values for fields in the list of resolveable conflicts (specified below), but are otherwise identical; keep one record and populate the conflicted fields with values as specified below in the list of resolveable conflicts:

List of resolveable conflicts:

receivedDate = the latest **receivedDate** among values entered,

testedDate = the latest **testedDate** among values entered,

testedBy = the latest **testedBy** among values entered,

deprecatedViallD = a pipe-concatenated list of all **deprecatedViallD** values from both records,

sampleCompromised = the **sampleCompromised** value that is not “No known compromise”;

For the record kept, **duplicatesInPathogenResultsQF** = 1; concatenate **remarks** as described above.

b. If multiple records still share identical **testingViallD**, **testNumber**, AND **testPathogenName** values after following step a (above) and have conflicting values or sample information for any field not specified in part a (i.e., the list of resolveable conflicts), then these records are unresolveable. In that case, retain both records and assign both with **duplicatesInPathogenResultsQF** = 2

6. No check for partial duplicates will be performed on mos_sorting_in, mos_pathogenpooling_in, or mos_archivepooling_in, because partial duplicates are expected to be resolved at data ingest.

6.7 Processing steps to run on mos_trapping_in

1. Rename fields used across multiple tables for clarity

Several tables share header names (i.e., remarks, laboratoryName), which will result in duplicated field names when the tables are merged. These fields need to be renamed accordingly in the L1 data.

Table 4: Partial input trapping data before renaming

plotID	collectDateTime	remarks
OSBS_056	20140409.1725	QA passed
OSBS_057	20140409.0940	
OSBS_057	20140409.1738	

mos_trapping_in:**remarks** populates mos_identification_pub:**trappingRemarks** and mos_samplingeffort_pub:**trappingRemarks**

mos_trapping_in:**recordedBy** populates mos_identification_pub:**trapRecordedBy** and mos_samplingeffort_pub:**trapRecordedBy**

mos_trapping_in:**sampleCompromised** populates mos_identification_pub:**sampleCompromisedAtTrapping** and mos_samplingeffort_pub:**sampleCompromisedAtTrapping**

Table 5: Partial output trapping data after renaming

plotID	collectDateTime	trappingRemarks
OSBS_056	20140409.1725	QA passed
OSBS_057	20140409.0940	
OSBS_057	20140409.1738	

2. Add spatial metadata to mos_trapping_in from the TOS spatial lookup table.

Table 6: Partial input trapping data before adding spatial data

plotID	collectDateTime
OSBS_056	20140409.0920
OSBS_056	20140409.1725
OSBS_056	20140410.1103

Match TOS spatial data:**plotID** to mos_trapping_in:**plotID** and add fields mos_trapping_in:**nlcdClass**, **decimalLatitude**, **decimalLongitude**, **geodeticDatum**, **coordinateUncertainty**, **elevation**, **elevationUncertainty** from the TOS spatial data lookup table.

Table 7: Partial output trapping data after adding spatial data

domainID	siteID	plotID	nlcdClass	collectDateTime
D03	OSBS	OSBS_056	evergreenForest	20140409.0920
D03	OSBS	OSBS_056	evergreenForest	20140409.1725
D03	OSBS	OSBS_056	evergreenForest	20140410.1103

3. Add field mos_trapping_in:**eventID** that will be used to identify unique bouts across years, based on site ID, year, and two-digit week number.

The eventID aggregates sampling events occurring over multiple continuous days into bouts. All samples within a bout are assigned a week number corresponding to the latest day of sampling.

Table 8: Partial input trapping data before adding eventID

siteID	collectDateTime	temp_year	temp_week
OSBS	20140409.0920	2014	14
OSBS	20140409.0940	2014	14
OSBS	20140409.1002	2014	14

Extract *temp_year* and *temp_week* based on Julian day (1-366) for each collectDateTime. *Example*: September 23, 2014 is the 266th Julian day of the year (non-leap year). For each site, identify sets of Julian days within 3 days of each other. Within these sets, change the lower values to match the highest value. *Example*: Sampling occurs on the following Julian days: [224, 266, 267]. The latter two days are fewer than three days apart and should be matched so that they are interpreted as being in the same week and therefore the same bout: [224, 267, 267]. Convert each Julian day value to a two digit week number (01-52) by dividing Julian day by 7 and rounding up. Any Julian days in excess of 364 are considered part of week 52.

siteID.*temp_year*.*temp_week* -> **eventID**

Table 9: Partial output trapping data after adding eventID

siteID	collectDateTime	temp_year	temp_week	eventID
OSBS	20140409.0920	2014	14	OSBS.2014.14
OSBS	20140409.0940	2014	14	OSBS.2014.14
OSBS	20140409.1002	2014	14	OSBS.2014.14

4. Add field mos_trapping_in:**trapCompromisedQF** that indicates any compromise in mosquito trap operation.

Table 10: Partial input trapping data before adding trapCompromised

collectDateTime	fanStatus	catchCupStatus	dryIceStatus	plotID
20140409.1016	Off	OK	Present	OSBS_054
20140410.1034	On		Present	OSBS_055
20140409.0920	On	OK	Present	OSBS_056
20140409.1002	On	OK	Present	OSBS_055

IF (**fanStatus** = 'On' AND **catchCupStatus** = 'OK' AND **dryIceStatus** = 'Present'), then **trapCompromisedQF** = 0

ELSE IF (**fanStatus** = 'Off' AND/OR (**catchCupStatus** = 'Disturbed' OR 'Missing') AND/OR **dryIceStatus** = 'Absent'), then **trapCompromisedQF** = 1

ELSE IF (**fanStatus** = NULL AND/OR **catchCupStatus** = NULL AND/OR **dryIceStatus** = NULL), then **trapCompromisedQF** = -1 (not evaluated)

Table 11: Partial output trapping data after adding trapCompromised

plotID	collectDateTime	fanStatus	catchCupStatus	dryIceStatus	trapCompromisedQF
OSBS_054	20140409.1016	Off	OK	Present	1
OSBS_055	20140410.1034	On		Present	-1
OSBS_056	20140409.0920	On	OK	Present	0
OSBS_055	20140409.1002	On	OK	Present	0

5. Add field mos_trapping_in:**nightOrDay** to indicate if collection occurred at night or during the day.

Table 12: Partial input trapping data before adding nightOrDay

sampleID	collectDateTime
OSBS_056.20140409.0920	20140409.0920
OSBS_057.20140409.0940	20140409.0940

IF mos_trapping_in:**collectDateTime** falls before noon (12am to 11:59am in local time; daylight savings may apply), then mos_trapping_in:**nightOrDay** = 'night'

ELSE IF mos_trapping_in:**collectDateTime** falls at or after noon (12pm to 11:59pm in local time; daylight savings may apply), then mos_trapping_in:**nightOrDay** = 'day'

Table 13: Partial output trapping data after adding nightOrDay

sampleID	collectDateTime	nightOrDay
OSBS_056.20140409.0920	20140409.0920	night
OSBS_057.20140409.0940	20140409.0940	night

6. Add field mos_trapping_in:**trapHours**, the length of time for which sampling occurred.

Table 14: Partial input trapping data before adding trapHours

sampleID	setDateTime	collectDateTime
OSBS_056.20140409.0920	20140408.1846	20140409.0920

sampleID	setDateTime	collectDateTime
OSBS_057.20140409.0940	20140408.1900	20140409.0940

Calculate the total hours between when the trap was set and collected: subtract mos_trapping_in:**setDateTime** from mos_trapping_in:**collectDateTime** and round the result to the nearest half-hour.

Table 15: Partial output trapping data after adding trapHours

sampleID	setDateTime	collectDateTime	trapHours
OSBS_056.20140409.0920	20140408.1846	20140409.0920	14.5
OSBS_057.20140409.0940	20140408.1900	20140409.0940	14.5

7. Truncate email address in mos_trapping_in:**recordedBy** to just the username (i.e., remove the @ symbol and the domain).

6.8 Processing steps to run on mos_sorting_in

1. Rename fields used across multiple tables for clarity

Table 16: Partial input sorting data before renaming

sampleID	totalWeight	sampleCompromised
OSBS_061.20140410.0958	2.85	No known compromise
OSBS_053.20140409.1055	1.4475	No known compromise
OSBS_054.20140409.1816	2.845	No known compromise

mos_sorting_in:**remarks** populates mos_identification_pub:**sortingRemarks**

mos_sorting_in:**laboratoryName** populates mos_identification_pub:**sortingLaboratoryName**

mos_sorting_in:**sampleCompromised** populates mos_identification_pub:**sampleCompromisedAtSorting** and mos_pathogenresults_pub:**sampleCompromisedAtSorting**

Table 17: Partial output sorting data after renaming

sampleID	totalWeight	sampleCompromisedAtSorting
OSBS_061.20140410.0958	2.85	No known compromise
OSBS_053.20140409.1055	1.4475	No known compromise

sampleID	totalWeight	sampleCompromisedAtSorting
OSBS_054.20140409.1816	2.845	No known compromise

6.9 Processing steps to run on mos_identification_in

1. Rename fields used across multiple tables for clarity

Table 18: Partial input identification data before renaming

subsampleID	laboratoryName	remarks
OSBS_053.20140409.1055.S.01	Colorado Mosquito Control	
OSBS_053.20140409.1055.S.01	Colorado Mosquito Control	

mos_identification_in:**remarks** populates mos_identification_pub:**identificationLabRemarks**.

mos_identification_in:**laboratoryName** populates mos_identification_pub:**idLaboratoryName**

mos_identification_in:**sampleCompromised** populates mos_identification_pub:**sampleCompromisedAtIDLab** and mos_pathogenresults_pub:**sampleCompromisedAtIDLab**.

Table 19: Partial output identification data after renaming

subsampleID	idLaboratoryName	identificationLabRemarks
OSBS_053.20140409.1055.S.01	Colorado Mosquito Control	
OSBS_053.20140409.1055.S.01	Colorado Mosquito Control	

2. Add required taxonomic fields mos_identification_in:**taxonID**, **taxonRank**, based on **scientificName**

Table 20: Partial identification data before adding taxonomic fields

subsampleID	scientificName
OSBS_052.20140409.1025.S.01	Culex fitchii
OSBS_052.20140409.1025.S.01	Culex fitchii
OSBS_052.20140409.1025.S.01	Culex nigripalpus
OSBS_052.20140410.1005.S.01	Culex nigripalpus
OSBS_053.20140410.1011.S.01	Culex nigripalpus

Table 21: Partial taxonTable used to inform taxonomic fields

taxonID	acceptedTaxonID	scientificName	taxonRank
AEDFIT	AEDFIT	Aedes fitchii	species
CULNIG	CULNIG	Culex nigripalpus	species
CULFIT	AEDFIT	Culex fitchii	species

The taxonTable is the lookup table called mos_names_status_list that contains all the taxonomic and range information for each taxon known to occur in the NEON realm.

Match mos_identification_in:**scientificName** with taxonTable:**scientificName** and find the corresponding taxonTable:*acceptedTaxonID* for that record (i.e., if mos_identification_in:**scientificName** is *Aedes fitchii*, then the taxonTable:*acceptedTaxonID* for that **scientificName** is AEDFIT). If taxonTable:*acceptedTaxonID* and taxonTable:**taxonID** are the same in the taxonTable (e.g., *Aedes fitchii* in Table 21), then use the information in that record of the taxonTable to populate mos_identification_in. If taxonTable:*acceptedTaxonID* and taxonTable:**taxonID** differ (e.g., *Culex fitchii* in Table 21), then lookup the record in the taxonTable where the taxonTable:**taxonID** is the same as the *acceptedTaxonID* in the previous step (e.g., the taxonTable:*acceptedTaxonID* for *Culex fitchii* is AEDFIT and the taxonTable:*acceptedTaxonID* and taxonTable:**taxonID** differ; find the taxonTable:**taxonID** entry where the taxonTable:**taxonID** matches AEDFIT). Use that record to populate the following entries in mos_identification_in.

IF a taxonTable:**taxonID** is found that matches the initial taxonTable:*acceptedTaxonID*, then:

1. populate mos_identification_in:**scientificName** with the taxonTable:**scientificName** for that taxonTable:**taxonID** where taxonTable:**taxonID** matches taxonTable:*acceptedTaxonID*
2. populate mos_identification_in:**taxonID** with taxonTable:**taxonID** for that taxonTable:**taxonID** where taxonTable:**taxonID** matches taxonTable:*acceptedTaxonID*
3. populate mos_identification_in:**taxonRank** with taxonTable:**taxonRank** for that taxonTable:**taxonID** where taxonTable:**taxonID** matches taxonTable:*acceptedTaxonID*

IF mos_identification_in:**targetTaxaPresent** = 'N' (e.g., no mosquitoes were present from that sample):

1. mos_identification_in:**scientificName** is NULL
2. mos_identification_in:**taxonID** is NULL
3. mos_identification_in:**taxonRank** is NULL

Table 22: Partial identification data after adding taxonomic fields

subsampleID	scientificName	taxonID	acceptedTaxonID	taxonRank
OSBS_052.20140409.1025.S.01	Aedes fitchii	AEDFIT	AEDFIT	species
OSBS_052.20140409.1025.S.01	Aedes fitchii	AEDFIT	AEDFIT	species
OSBS_052.20140409.1025.S.01	Culex nigripalpus	CULNIG	CULNIG	species
OSBS_052.20140410.1005.S.01	Culex nigripalpus	CULNIG	CULNIG	species
OSBS_053.20140410.1011.S.01	Culex nigripalpus	CULNIG	CULNIG	species

3. Generate mos_identification_in:**idqQF** flag

Table 23: Partial input identification data before adding scientificName qualifier and flag

subsampleID	scientificName	identificationQualifier	taxonRank
OSBS_058.20140410.0922.S.01	Aedes hendersoni	cf. species	species
OSBS_057.20140827.0720.S.01	Aedes sp.	cf. species	genus
OSBS_054.20140909.1745.S.01	Aedes hendersoni	cf. genus	species
OSBS_054.20140909.1745.S.01	Coquillettidia perturbans		species

The **idqQF** quality flag is intended to evaluate if the **identificationQualifier** field matches a lookup table value AND **taxonRank** is at the same level as the **identificationQualifier**. An **identificationQualifier** is on the same level as **taxonRank** when they describe the same **taxonRank** (i.e., the **identificationQualifier** "cf. species" contains the word "species", this is on the same level as **taxonRank** "species"). The taxonomic hierachy proceeds from higher taxonomic ranks to lower taxonomic ranks in the following order: family, subfamily, tribe, genus, species, subspecies. A **taxonRank** is higher than **identificationQualifier** if it describes a taxonomic rank at a higher level than that **identificationQualifier** (i.e, **taxonRank** = "genus", but **identificationQualifier** = "cf. species"). Conversely, **taxonRank** is lower than **identificationQualifier** if it describes a taxonomic rank at a lower level than that **identificationQualifier** (i.e, **taxonRank** = "species", but **identificationQualifier** = "cf. genus").

Create a quality flag called **idqQF** that has the following values:

-1 = test not run (no idq code provided or **taxonRank** is "speciesGroup")

0 = idq code exists in the lookup table and its rank matched taxonRank

1 = idq code rank is lower than taxonRank, no idq code needed even though it was provided

2 = idq code rank is higher than taxonRank, issue couldn't be rectified OR taxonRank did not match for the taxonID and acceptedTaxonID (in the taxon table).

Special flagging rules are needed for newly discovered species. IF **identificationQualifier** code is "sp. nr.":

a. Flag the record if taxonRank is not species or subspecies; **idqQF** = 2

b. Otherwise **idqQF** = 0

IF **identificationQualifier** code is “spp. nov.” or “sp. nov.”:

a. Flag the record if taxonRank is not genus or family; **idqQF** = 2

b. Otherwise **idqQF** = 0

Table 24: Partial output identification data after adding scientificName qualifier and flag

subsampleID	scientificName	identificationQualifier	taxonRank	idqQF
OSBS_058.20140410.0922.S.01	Aedes hendersoni	cf. species	species	0
OSBS_057.20140827.0720.S.01	Aedes sp.	cf. species	genus	1
OSBS_054.20140909.1745.S.01	Aedes hendersoni	cf. genus	species	2
OSBS_054.20140909.1745.S.01	Coquillettidia perturbans		species	-1

6.10 Processing steps to run on mos_pathogenresults_in

1. Rename fields used across multiple tables for clarity

Table 25: Partial input pathogen testing data before renaming

testingViallD	laboratoryName	sampleCompromised
OSBS.2014.20.COQPER.F.T.12	Connecticut Agricultural Experiment Station	No known compromise
OSBS.2014.20.COQPER.F.T.13	Connecticut Agricultural Experiment Station	No known compromise

mos_pathogenresults_in:**sampleCompromised** populates mos_pathogenresults_pub:**sampleCompromisedAtTesting**

Table 26: Partial output pathogen testing data after renaming

testingViallD	laboratoryName	sampleCompromisedAtTesting
OSBS.2014.20.COQPER.F.T.12	Connecticut Agricultural Experiment Station	No known compromise
OSBS.2014.20.COQPER.F.T.13	Connecticut Agricultural Experiment Station	No known compromise

6.11 Processing steps to generate mos_samplingeffort_pub

This table provides data on field sampling effort and whether or not mosquitoes were present in each trap, with confirmation from sorting and identification data.

1. Add field mos_samplingeffort_pub:**sampleLostQF** to track presence of samples and affiliated subsamples from mos_trapping_in, through mos_sorting_in and mos_identification_in.

Table 27: Partial input mos_samplingeffort_pub before adding sampleLostQF

sampleID	targetTaxaPresent	collectDateTime
OSBS_060.20140423.0829	Y	20140423.0829
OSBS_053.20140423.0906	Y	20140423.0906
OSBS_052.20140423.0913	Y	20140423.0913
OSBS_059.20140603.0840	Y	20140603.0840

Table 28: Partial sorting data used to inform sampleLostQF

sampleID	subsampleID	subsampleWeight	bycatchWeight
OSBS_052.20140423.0913	OSBS_052.20140423.0913.S.01	1.0025	0
OSBS_053.20140423.0906	OSBS_053.20140423.0906.S.01	1.255	0
OSBS_060.20140423.0829	OSBS_060.20140423.0829.S.01	1.1825	0

Table 29: Partial identification data used to inform sampleLostQF

subsampleID	targetTaxaPresent	individualCount	scientificName
OSBS_052.20140423.0913.S.01	Y	12	Aedes sp.
OSBS_052.20140423.0913.S.01	Y	83	Anopheles crucians
OSBS_059.20140603.0840.S.01	Y	4	Culex salinarius
OSBS_059.20140603.0840.S.01	Y	50	Culex sp.

Samples can be lost in either the sorting or identification stage.

IF mos_trapping_in:**targetTaxaPresent** == 'Y' AND:

Lost before sorting: IF mos_trapping_in:**sampleID** is not in mos_sorting_in:**sampleID**, THEN mos_samplingeffort_pub:**sampleLostQF** = 1

Records from sorting with mosquitoes present that do not appear in the identification data should be considered lost.

Lost before identification: IF *sortTTP* ==‘Y’ AND its associated mos_sorting_in:**subsampleID** is NOT in mos_identification_in:**subsample** THEN mos_samplingeffort_pub:**sampleLostQF** = 1

Otherwise, if subsamples can be traced through mos_sorting_in and mos_identification_in then the sample is not considered lost.

ELSE IF *sortTTP* ==‘Y’ AND mos_sorting_in:**subsampleID** is in mos_identification_in:**subsampleID**, THEN mos_samplingeffort_pub:**sampleLostQF** = 0

Table 30: Partial output mos_samplingeffort_pub after adding sampleLostQF

sampleID	targetTaxaPresent	collectDateTime	sampleLostQF
OSBS_060.20140423.0829	Y	20140423.0829	0
OSBS_053.20140423.0906	Y	20140423.0906	0
OSBS_052.20140423.0913	Y	20140423.0913	0
OSBS_059.20140603.0840	Y	20140603.0840	0

2. mos_samplingeffort_pub:**targetTaxaPresent** will reflect a consensus among field collection, sorting, and taxonomic identification. Taxonomist-determined values are prioritized first, followed by sorting values indicating no mosquitoes, otherwise the values from trapping data are used.

    a. For each record of *mos_trapping_in* where **sampleID** is not NULL:
    b. Find the set of records in *mos_sorting_in* where *mos_sorting_in*:**sampleID** = *mos_trapping_in*:**sampleID** from step (a)

        i. If no records exist where *mos_sorting_in*:**sampleID** = *mos_trapping_in*:**sampleID**, then *mos_samplingeffort_pub*:**targetTaxaPresent** = *mos_trapping_in*:**targetTaxaPresent**
        ii. Else: find the set of records in *mos_identification_in* where *mos_identification_in*:**subsampleID** is contained within the set of *mos_sorting_in*:**subsampleID** values associated with the records from step b (above).
        iii. If no records exist where *mos_identification_in*:**subsampleID** is contained within the set of *mos_sorting_in*:**subsampleID** values in the records from step b (above), then *mos_samplingeffort_pub*:**targetTaxaPre** = *mos_trapping_in*:**targetTaxaPresent**
        iv. Else: if the set of records in step b.ii contains at least 1 record where *mos_identification_in*:**targetTaxaPresent** = ‘Y’, *mos_samplingeffort_pub*:**targetTaxaPresent** = Y; if all records contain ‘N’ or NULL *mos_samplingeffort_pub*:**targetTaxaPresent** = N

    c. If *mos_trapping_in*:**sampleID** is NULL: *mos_samplingeffort_pub*:**targetTaxaPresent** = *mos_trapping_in*:**targetTaxaPresent**

Table 31: Partial input mos_samplingeffort_pub before updating targetTaxaPresent

sampleID	targetTaxaPresent	collectDateTime
OSBS_059.20140409.1043	Y	20140409.1043
OSBS_059.20140409.1841	Y	20140409.1841
OSBS_060.20140409.1856	Y	20140409.1856
OSBS_061.20140410.0958	Y	20140410.0958

Table 32: Partial sorting data (mos_sorting_in) used to inform targetTaxaPresent

subsampleID	subsampleWeight	bycatchWeight
OSBS_059.20140409.1043.S.01	1.9075	0.4
OSBS_060.20140409.1856.S.01	3	3
OSBS_061.20140410.0958.S.01	2.85	0.78
OSBS_059.20140409.1841.S.01	1.9075	1.22

Table 33: Partial identification data (mos_identification_in) used to inform targetTaxaPresent

subsampleID	targetTaxaPresent	individualCount	scientificName
OSBS_059.20140409.1043.S.01	Y	1	Aedes mitchellae
OSBS_059.20140409.1043.S.01	Y	50	Anopheles crucians
OSBS_060.20140409.1856.S.01	N	0	
OSBS_061.20140410.0958.S.01	Y	38	Anopheles crucians
OSBS_061.20140410.0958.S.01	Y	1	Coquillettidia perturbans
OSBS_059.20140409.1841.S.01	Y	3	Anopheles crucians
OSBS_059.20140409.1841.S.01	Y	1	Culex sp.

Table 34: Partial output mos_samplingeffort_pub after updating targetTaxaPresent

sampleID	targetTaxaPresent	collectDateTime
OSBS_059.20140409.1043	Y	20140409.1043
OSBS_059.20140409.1841	Y	20140409.1841
	N	20140409.1856
OSBS_061.20140410.0958	Y	20140410.0958

6.12 Processing steps to generate mos_identification_pub

mos_identification_pub is filtered to only contain records with expert-confirmed presence of mosquitoes. All other records, including those with incomplete information, will still appear in mos_samplingeffort_pub.

1. Create mos_identification_pub.

Table 35: Partial input trapping data to be merged

plotID	collectDateTime	sampleID
OSBS_057	20140409.1738	OSBS_057.20140409.1738
OSBS_054	20140409.1816	OSBS_054.20140409.1816

Table 36: Partial input sorting data to be merged

subsampleID	sampleID	totalWeight
OSBS_052.20140409.1025.S.01	OSBS_052.20140409.1025	2.49
OSBS_052.20140410.1005.S.01	OSBS_052.20140410.1005	2.49

Table 37: Partial input table identification data to be merged

subsampleID	scientificName	individualCount
OSBS_052.20140409.1025.S.01	Anopheles crucians	106
OSBS_052.20140410.1005.S.01	Coquillettidia perturbans	10

Partially join mos_identification_in:**subsampleID** with mos_sorting_in:**subsampleID** to generate mos_identification_pub (keep *all* records from the mos_identification_in, but only keep records from mos_sorting_in that match on **subsampleID**)

Partially join the resulting mos_identification_pub:**sampleID** with mos_trapping_in:**sampleID** to complete mos_identification_pub (keep *all* records from the mos_identification_pub, but only keep records from mos_trapping_in that match on **sampleID**)

Table 38: Partial output mos_identification_pub after merging inputs

collectDateTime	sampleID	totalWeight	scientificName	individualCount
20140409.1025	OSBS_052.20140409.1025	2.49	Anopheles crucians	106

collectDateTime	sampleID	totalWeight	scientificName	individualCount
20140410.1005	OSBS_052.20140410.1005	2.49	Coquillettidia perturbans	10

2. Remove records for which mos_identification_pub:**targetTaxaPresent** is 'N'.

3. Add fields mos_identification_pub:**percentCounted** [Equation 1], the percentage of the sample that was identified, and **estimatedAbundance**, the total estimated number of individuals of this species and sex in this trap [Equation 2].

Table 39: Partial input mos_identification_pub before adding estimatedAbundance

subsampleID	totalWeight	subsampleWeight	scientificName	individualCount
OSBS_059.20140410.0952.S.01	1.9075	1.9075	Culiseta melanura	32
OSBS_060.20140409.1115.S.01	6	3	Anopheles crucians	230
OSBS_060.20140409.1115.S.01	6	3	Anopheles sp.	5

percentCounted is the percentage of the total sample that the subsample comprises:

$$percentCounted = \frac{\mathbf{subsampleWeight}}{\mathbf{totalWeight}} * 100 \quad (1)$$

Using mos_identification_pub:**subsampleWeight**, **totalWeight**, and **individualCount**, estimate the total number (**estimatedAbundance**) of individuals, rounded to the nearest integer:

$$estimatedAbundance = \mathbf{individualCount} * \frac{\mathbf{totalWeight}}{\mathbf{subsampleWeight}} \quad (2)$$

If ANY of **subsampleWeight**, **bycatchWeight**, **totalWeight**, **individualCount** are NULL: **estimatedAbundance** is not calculated for that record.

Table 40: Partial output mos_identification_pub after adding estimatedAbundance

subsampleID	scientificName	estimatedAbundance	percentCounted
OSBS_059.20140410.0952.S.01	Culiseta melanura	32	100
OSBS_060.20140409.1115.S.01	Anopheles crucians	460	50
OSBS_060.20140409.1115.S.01	Anopheles sp.	10	50

4. Add field mos_identification_pub:**taxonRangeQF** where taxa appear to be out of range.

Table 41: Partial input mos_identification_pub before adding taxonRangeQF

subsampleID	domainID	scientificName	taxonRank
OSBS_052.20140409.1025.S.01	D03	Aedes fitchii	species
OSBS_057.20140827.1747.S.01	D03	Aedes albopictus	species
OSBS_057.20140409.0940.S.01	D03	Aedes atlanticus	species

Table 42: Partial input taxonomic data used to inform taxonRangeQF

scientificName	d03NativeStatusCode
Aedes atlanticus	N
Aedes fitchii	A
Aedes albopictus	I

For each record, use mos_identification_pub:**domainID** determine which taxonTable field to use as reference, *NativeStatusCode*. Match mos_identification_pub:**scientificName** to taxonTable:**scientificName**, and get the corresponding value from the appropriate reference field: presumed absent ('A'), a known native species ('N'), a known introduced species ('I'), or a taxon with unknown distribution ('UNK').

IF taxonTable[**scientificName**,*NativeStatusCode*] == 'A': mos_identification_pub:**taxonRangeQF** = 1

ELSE IF taxonTable[**scientificName**,*NativeStatusCode*] == ('N'|'I'|'UNK'): mos_identification_pub:**taxonRangeQF** = 0

Table 43: Partial output mos_identification_pub after adding taxonRangeQF

subsampleID	scientificName	taxonRangeQF
OSBS_052.20140409.1025.S.01	Aedes fitchii	1
OSBS_057.20140827.1747.S.01	Aedes albopictus	0
OSBS_057.20140409.0940.S.01	Aedes atlanticus	0

6.13 Processing steps to generate mos_pathogenresults_pub

1. Modify testPathogenName if necessary.

pathogenTaxonTable is mpt_names_status_list.

Match mos_pathogenresults_in:**testPathogenName** to pathogenTaxonTable:**scientificName**, to get the corresponding pathogenTaxonTable:*acceptedTaxonID*.

Populate mos_pathogenresults_in:**testPathogenName** with the pathogenTaxonTable:**scientificName** where pathogenTaxonTable:**taxonID** is equal to the *acceptedTaxonID* found in the previous step.

2. Merge inputs to create mos_pathogenresults_pub.

Table 44: Partial input trapping data from mos_trapping_in to be merged

siteID	sampleID	nlcdClass
OSBS	OSBS_054.20140702.0905	evergreenForest
OSBS	OSBS_052.20140702.0913	evergreenForest
OSBS	OSBS_053.20140702.0920	evergreenForest

Table 45: Partial input sorting data from mos_sorting_in to be merged

subsampleID	sampleID
OSBS_052.20140702.0913.S.01	OSBS_052.20140702.0913
OSBS_053.20140702.0920.S.01	OSBS_053.20140702.0920
OSBS_054.20140702.0905.S.01	OSBS_054.20140702.0905

Table 46: Partial input identification data from mos_identification_in to be merged

subsampleID	scientificName	testingID	individualCount
OSBS_052.20140702.0913.S.01	Coquillettidia perturbans	OSBS.2014.26.COQPER.F.T	206
OSBS_052.20140702.0913.S.01	Culex erraticus	OSBS.2014.26.CULERR.F.T	41
OSBS_053.20140702.0920.S.01	Coquillettidia perturbans	OSBS.2014.26.COQPER.F.T	220
OSBS_053.20140702.0920.S.01	Culex erraticus	OSBS.2014.26.CULERR.F.T	67
OSBS_054.20140702.0905.S.01	Coquillettidia perturbans	OSBS.2014.26.COQPER.F.T	31
OSBS_054.20140702.0905.S.01	Culex erraticus	OSBS.2014.26.CULERR.F.T	242
OSBS_054.20140702.0905.S.01	Culex nigripalpus	OSBS.2014.26.CULNIG.F.T	5

Table 47: Partial input pathogen pooling data from mos_pathogenpooling_in to be merged

testingVialID	testingID	poolSize
OSBS.2014.26.COQPER.F.T.01	OSBS.2014.26.COQPER.F.T	50
OSBS.2014.26.CULERR.F.T.01	OSBS.2014.26.CULERR.F.T	50

testingVialID	testingID	poolSize
OSBS.2014.26.CULNIG.F.T.01	OSBS.2014.26.CULNIG.F.T	35

Table 48: Partial input pathogen results data from mos_pathogenresults_in to be merged

testingVialID	testResult	testPathogenName
OSBS.2014.26.COQPER.F.T.01	Negative	virus
OSBS.2014.26.CULERR.F.T.01	Negative	virus
OSBS.2014.26.CULNIG.F.T.01	Negative	virus

Partially join mos_pathogenresults_in:**testingVialID** with mos_pathogenpooling_in:**testingVialID** to generate mos_pathogenresults_pub (keep *all* records from the mos_pathogenresults_in, but only keep records from mos_pathogenpooling_in that match on **testingVialID**)

Add fields mos_pathogenresults_pub:**scientificName**, **sex**, **taxonRank**, **identificationQualifier**, and **taxonID**, based on a partial join of mos_pathogenresults_pub:**testingID** with mos_identification_in:**testingID** Keep *all* records from the original mos_pathogenresults_pub, but only keep records from mos_identification_in that match on **testingID**.

Add fields mos_pathogenresults_pub: the **decimalLatitude**, **decimalLongitude**, **geodeticDatum**, **coordinateUncertainty**, **elevation**, **elevationUncertainty**, **samplingProtocolVersion**, **sampleTiming**, and **eventID**, based on a partial join of mos_pathogenresults_pub:**sampleID** to mos_trapping_in:**sampleID**. Keep all records in mos_pathogenresults_pub but only keep records from mos_trapping_in that match on **sampleID**.

Numerical site-level statistics can be determined by reporting a mean value from the TOS Spatial Database for **decimalLatitude**, **decimalLongitude**, and **elevation** across all plots that *contributed to the* **testingVialID** and include mos in the applicable module field. Uncertainties (e.g. **coordinateUncertainty** and **elevationUncertainty**) should be calculated by determining location of the plot furthest from the reported site-level centroid (using **easting**, **northing**, and **elevation** in the TOS Spatial Data) and reporting the distance between the centroid and the furthest plot from the centroid at the site including **coordinateUncertanty** associated with that **plotID**.

Example: A centroid is reported at Latitude = 44.05, Longitude = -71.26. The furthest plot from that centroid at that site is 100 meters away and has an **coordinateUncertainty** of 5 meters. The site-level **coordinateUncertainty** is 105 meters. Likewise, **elevationUncertainty** would be calculated by finding the greatest difference in **elevation** from the mean elevation at a site and adding the **elevationUncertainty** associated with that plot (e.g., if mean **elevation** is 34 meters above sea level, the greatest difference in **elevation** from the mean is 14 meters and the **elevationUncertainty** associated with that plot is 0.5 meters: then the site-level **elevationUncertainty** is 14.5 meters).

Table 49: Partial output mos_pathogenResults_pub after merging inputs

testingVialID	poolSize	scientificName	sex	testResult
OSBS.2014.26.COQPER.F.T.01	50	Coquillettidia perturbans	F	Negative
OSBS.2014.26.CULERR.F.T.01	50	Culex erraticus	F	Negative

testingVialID	poolSize	scientificName	sex	testResult
OSBS.2014.26.CULNIG.F.T.01	35	Culex nigripalpus	F	Negative

3. Add field mos_pathogenresults_pub: **collectDate**

Table 50: Partial input mos_pathogenResults_pub before adding startCollectDate and endCollectDate

eventID	testingVialID	poolSize	scientificName
OSBS.2014.26	OSBS.2014.26.COQPER.F.T.01	50	Coquillettidia perturbans
OSBS.2014.26	OSBS.2014.26.CULERR.F.T.01	50	Culex erraticus
OSBS.2014.26	OSBS.2014.26.CULNIG.F.T.01	35	Culex nigripalpus

Table 51: Partial input trapping data from mos_trapping_in used to inform startCollectDate and endCollectDate

eventID	collectDateTime
OSBS.2014.26	20140701.0756
OSBS.2014.26	20140701.0805
OSBS.2014.26	20140702.0959
OSBS.2014.26	20140702.1010

Populate **startCollectDate** and **endCollectDate** with the earliest and latest date (respectively) associated with a particular **eventID** (i.e., if an **eventID** is associated with both 20140731.1000 and 20140801.0900, **startCollectDate** = 20140731.1000 and **endCollectDate** = 20140801.0900).

Table 52: Partial output mos_pathogenResults_pub after adding startCollectDate and endCollectDate

eventID	testingVialID	poolSize	scientificName	startCollectDate	endCollectDate
OSBS.2014.26	OSBS.2014.26.COQPER.F.T.01	50	Coquillettidia perturbans	20140701.0756	20140702.1010
OSBS.2014.26	OSBS.2014.26.CULERR.F.T.01	50	Culex erraticus	20140701.0756	20140702.1010
OSBS.2014.26	OSBS.2014.26.CULNIG.F.T.01	35	Culex nigripalpus	20140701.0756	20140702.1010

4. Add field mos_pathogenresults_pub:**nonStandardPoolQF**, an indicator that individuals were pooled in a manner inconsistent with the standard protocol (i.e. different bouts or species were combined).

Table 53: Partial input mos_pathogenResults_pub before adding nonStandardPoolQF

testingID	testingVialID	scientificName	sex	startCollectDate
OSBS.2014.14.COQPER.F.T	OSBS.2014.14.COQPER.F.T.01	Aedes fitchii	F	20140409.0920
OSBS.2014.14.COQPER.F.T	OSBS.2014.14.COQPER.F.T.01	Coquillettidia perturbans	F	20140409.0920
OSBS.2014.14.CULSAL.F.T	OSBS.2014.14.CULSAL.F.T.01	Culex salinarius	F	20140409.0920

If samples are pooled in a way that deviates from the species/sex/bout combination expected, then a flag will be generated.

Samples with a **testingID** that correspond to a single mosquito species/sex/bout combination:

nonStandardPoolQF = 0

Any samples that share a **testingID** but differ in their **scientificName**, **sex**, **taxonID** or **taxonRank** within the same bout:

nonStandardPoolQF = 1,

scientificName = “Culicidae sp.”,

sex = “U”,

identificationQualifier = NULL,

taxonRank = “family”,

taxonID = “CULSPP”.

Table 54: Partial output mos_pathogenResults_pub after adding nonStandardPoolQF

testingVialID	scientificName	sex	startCollectDate	nonStandardPoolQF
OSBS.2014.14.COQPER.F.T.01	Culicidae sp.	U	20140409.0920	1
OSBS.2014.14.COQPER.F.T.01	Culicidae sp.	U	20140409.0920	1
OSBS.2014.14.CULSAL.F.T.01	Culex salinarius	F	20140409.0920	0

5. Add fields mos_pathogenresults_pub:**trapCompromisedQF**, **sampleCompromisedAtSortingQF**, and **sampleCompromisedAtIDLabQF**

Table 55: Partial input mos_pathogenResults_pub before adding compromised flags

siteID	startCollectDate	testingVialID	trapCompromisedQF
OSBS	20140701.0756	OSBS.2014.26.COQPER.F.T.01	0
OSBS	20140701.0756	OSBS.2014.26.COQPER.F.T.01	0

siteID	startCollectDate	testingVialID	trapCompromisedQF
OSBS	20140701.0756	OSBS.2014.26.COQPER.F.T.01	1

For all records with the same **testingID**:

1. IF ANY record's **trapCompromisedQF** is NOT 0

a. THEN mos_pathogenresults_pub:**trapCompromisedQF** = 1

2. IF ANY record's **sampleCompromisedAtTrapping** is NOT 'No known compromise'

a. THEN mos_pathogenresults_pub:**sampleCompromisedAtTrapping** is a pipe-separated list of all the unique entries from mos_trapping_in:**sampleCompromised** that are a parent to that **testingID** (sort entries in the list alphabetically), excluding 'No known compromise'
b. ELSE mos_pathogenresults_pub:**sampleCompromisedAtTrapping** is 'No known compromise'

3. IF ANY record's **sampleCompromisedAtSorting** is NOT 'No known compromise'

a. THEN mos_pathogenresults_pub:**sampleCompromisedAtSorting** is a pipe-separated list of all the unique entries from mos_sorting_in:**sampleCompromised** that are a parent to that **testingID** (sort entries in the list alphabetically), excluding 'No known compromise'
b. ELSE mos_pathogenresults_pub:**sampleCompromisedAtSorting** is 'No known compromise'

4. IF ANY record's **sampleCompromisedAtIDLab** is NOT 'No known compromise'

a. THEN mos_pathogenresults_pub:**sampleCompromisedAtIDLab** is a pipe-separated list of all the unique entries from mos_identification_in:**sampleCompromised** that are a parent to that **testingID** (sort entries in the list alphabetically), excluding 'No known compromise'
b. ELSE mos_pathogenresults_pub:**sampleCompromisedAtIDLab** is 'No known compromise'

Table 56: Partial output mos_pathogenResults_pub after adding compromised flags

testingVialID	trapCompromisedQF	sampleCompromisedAtSorting
OSBS.2014.26.COQPER.F.T.01	1	No known compromise
OSBS.2014.26.COQPER.F.T.01	1	No known compromise
OSBS.2014.26.COQPER.F.T.01	1	No known compromise

6.14 Processing steps to generate mos_archival_pub

mos_archival_pub provides information on specimens that were archived.

1. Merge ingests to create mos_archival_pub.

Table 57: Partial input trapping data from mos_trapping_in to be merged

sampleID	collectDateTime
OSBS_055.20140409.1002	20140409.1002
OSBS_054.20140409.1016	20140409.1016
OSBS_052.20140409.1025	20140409.1025
OSBS_053.20140410.1011	20140410.1011
OSBS_054.20140410.1022	20140410.1022

Table 58: Partial input identification data from mos_identification_in to be merged

scientificName	sex	archiveID	subsampleID
Culex erraticus	F	OSBS.2014.14.CULERR.F.A	OSBS_052.20140409.1025.S.01
Culex erraticus	F	OSBS.2014.14.CULERR.F.A	OSBS_053.20140410.1011.S.01
Culex erraticus	F	OSBS.2014.14.CULERR.F.A	OSBS_054.20140409.1016.S.01
Culex erraticus	F	OSBS.2014.14.CULERR.F.A	OSBS_054.20140410.1022.S.01
Culex erraticus	F	OSBS.2014.14.CULERR.F.A	OSBS_055.20140409.1002.S.01

Table 59: Partial input archive pooling data from mos_archivepooling_in to be merged

archiveID	archiveVialIDList	archivedCount
OSBS.2014.14.CULERR.F.A	OSBS.2014.14.CULERR.F.A.01	165

Partially join mos_identification_in:**archiveID** with mos_archivepooling_in:**archiveID** -> mos_archival_pub. Keep all records from mos_archivepooling_in, keep only matching records from mos_identification_in.

Partially join mos_trapping_in:**eventID** with mos_archival_pub:**eventID**. Keep all records from mos_archival_pub, keep only matching records from mos_trapping_in.

Site level GIS data will be appended in the same way as in the mos_pathogenresults_pub

Lookup mos_archival_pub:**archiveID** within mos_identification_in. Populate the following identification informa-

tion within mos_archival_pub with values from mos_identification_in that match on **archiveID** such that:

mos_archival_pub:**scientificName** = mos_identification_in:**scientificName**,

mos_archival_pub:**sex** = mos_identification_in:**sex**,

mos_archival_pub:**identificationQualifier** = mos_identification_in:**identificationQualifier**,

mos_archival_pub:**taxonID** is the **acceptedTaxonID** from the taxon table associated with the assigned **scientificName**,

mos_archival_pub:**taxonRank**) is the **taxonRank** from the taxon table associated with the assigned **scientificName**.

IF more than one taxonomic value is associated with a particular mos_archival_pub:**archiveID** within the mos_identification_in data (i.e., two records in mos_identification_in have the same value for **archiveID** and different values for **scientificName**, **sex**, **taxonID**, or **taxonRank**), THEN all records with that mos_archival_pub:**archiveID** should be assigned the following values:

scientificName = “Culicidae sp.”,

sex = “U”,

identificationQualifier = NULL,

taxonID = “CULSPP”,

taxonRank = “family”.

If a particular mos_archival_pub:**archiveID** has multiple **identificationQualifier** values associated with it, but no differences in **scientificName**, **sex**, **taxonID** and **taxonRank**, then set **identificationQualifier** to any non-null existing value.

Table 60: Partial output mos_archival_pub after merging inputs

archiveID	scientificName	archivedCount
OSBS.2014.14.CULERR.F.A	Culex erraticus	165

2. Add fields mos_archival_pub: **startCollectDate** and mos_archival_pub: **endCollectDate**.

Table 61: Partial input mos_archival_pub before adding startCollectDate and endCollectDate

eventID	archiveVialIDList	scientificName	archivedCount
OSBS.2014.14	OSBS.2014.14.AEDMIT.F.A.01	Aedes mitchellae	11

Table 62: Partial input trapping table used to inform startCollectDate and endCollectDate

collectDateTime	eventID
20140409.1025	OSBS.2014.14
20140410.1005	OSBS.2014.14

Populate **startCollectDate** and **endCollectDate** with the earliest and latest **collectDateTime** (respectively) associated with a particular **eventID** (i.e., if an **eventID** is associated with both 20140731.1000 and 20140801.0900, **startCollectDate** = 20140731.1000 and **endCollectDate** = 20140801.0900).

Table 63: Partial output mos_archival_pub after adding startCollectDate and endCollectDate

startCollectDate	endCollectDate	eventID	archiveVialIDList	scientificName	archivedCount
20140409.0920	20140410.1103	OSBS.2014.14	OSBS.2014.14.AEDMIT.F.A.01	Aedes mitchellae	11

6.15 Steps to run for all publication tables

Only certain columns are needed for each publication table. All tables should only have the fields listed in the publication workbook.

7 Uncertainty

7.1 Analysis of Uncertainty

7.2 Reported Uncertainty

Although no quantitative algorithms are available to incorporate many of these sources of uncertainty into the associated data products, NEON can produce summary uncertainty reports for these observational data products. Assessing error rates in taxonomic identifications is beyond the scope of this document.

8 Scientific and Educational Applications

NEON mosquito and mosquito-borne pathogen sampling data will characterize variation in abundance, diversity, and phenology of mosquitoes and prevalence of mosquito-borne pathogens across sampling locations through time. Investigating and explaining this variation through analyses involving data on biotic and abiotic factors believed to influence infection dynamics will shed light on the demography of mosquito populations and epidemiology of the mosquito-borne pathogens tested for as part of this sampling.

9 Future Modifications and Plans

In the *mos_samplingeffort_pub*, the absence of a subsample is acceptable if no mosquitoes were present. In a future rev of this data product, for any sorting subsample that was all bycatch *mos_samplingeffort_pub*:**targetTaxaPresent** should be 'N', *mos_samplingeffort_pub*:**sampleID** should be NULL and *mos_samplingeffort_pub*:**sampleLostQF** should be 0. In the current implementation, samples sent by field operations for taxonomy that are entirely bycatch may be incorrectly flagged as a 'lost sample' instead of being recategorized as *mos_samplingeffort_pub*:**targetTaxaPresent** = 'N'. In 2014, ~0.6% of samples sent to the taxonomist were all bycatch.

Future additions to this ATBD will include activities related to DNA barcoding. Some mosquitoes that have been identified to species at external facilities will be returned to domain labs for pointing, photography and tissue extraction (leg removed) for DNA barcoding. Afterward, these specimens will be sent to an archive facility, while the photographs will be submitted along with tissue samples and associated collection data to be included with the barcode record.

Improvements in the software used to record data in the field would substantially improve data quality. This document describes some of the software parameters that will be in use on ruggedized modile devices to input trapping data in the field. However, future iterations of software for these mobile devices should compare GPS coordinates of the selected **plotID** with current GPS location detected by the device. Individuals should receive warnings if the plotID entered is not congruent with the current GPS location measured by the device within a tolerance of 100 meters.

10 Bibliography

AIBSnews. 2007. NEON design 2007. BioScience 57:198-200.

Bhatt, S., P. W. Gething, O. J. Brady, J. P. Messina, A. W. Farlow, C. L. Moyes, J. M. Drake, J. S. Brownstein, A. G. Hoen, O. Sankoh, M. F. Myers, D. B. George, T. Jaenisch, G. R. W. Wint, C. P. Simmons, T. W. Scott, J. J. Farrar, and S. I. Hay. 2013. The global distribution and burden of dengue. Nature 496:504-507.

Biggerstaff, B. 2005. PooledInfRate software. Vector-Borne and Zoonotic Diseases 5:420-421.

Brault, A. C. and W. K. Reisen. 2013. Environmental purturbations that influence arbovial host range: insights into emergence mechanisms. Pages 57-75 in S. K. Singh, editor. Viral Infections and Global Change. John Wiley and Sons, Inc., New York, NY.

Buckner, E. A., M. S. Blackmore, S. W. Golladay, and A. P. Covich. 2010. Weather and landscape factors associated with adult mosquito abundance in southwestern Georgia, U.S.A. Journal of Vector Ecology 36:269-278.

Caminade, C., J. M. Medlock, E. Ducheyne, K. M. McIntyre, S. Leach, M. Baylis, and A. P. Morse. 2012. Suitability of European climate for the Asian tiger mosquito Aedes albopictus: recent trends and future scenarios. Journal of the Royal Society Interface 9:2708-2717.

Daubney, R., J. R. Hudson, and P. C. Garnham. 1931. Enzootic hepatitis or Rift Valley fever. An undescribed virus disease of sheep cattle and man from East Africa. Journal of Pathology and Bacteriology 34:545-579.

Eisen, L. and C. G. Moore. 2013. Aedes (Stegomyia) aegypti in the continental United States: A vector at the cool margin of its geographic range. Journal of Medical Entomology 50:467-478.

Epstein, P. R., H. F. Diaz, S. Elias, G. Grabherr, N. E. Graham, W. J. M. Martens, E. Mosley-Thompson, and J. Susskind. 1998. Biological and physical signs of climate change: Focus on mosquito-borne diseases. Bulletin of the American Meteorological Society 79:409-417.

Gould, E. A. and S. Higgs. 2009. Impact of climate change and other factors on emerging arbovirus diseases. Transactions of the Royal Society of Tropical Medicine and Hygiene 103:109-121.

Harrigan, R. J., H. A. Thomassen, W. Buermann, and T. B. Smith. 2014. A continental risk assessment of West Nile virus under climate change. Global Change Biology.

Kissling, R. E., R. W. Chamberlain, M. E. Eidson, R. K. Sikes, and M. A. Bucca. 1954. Studies on the North American arthropod-borne encephalitides. 2. Eastern Equine Encephalites in Horses. American Journal of Hygiene 60:237-250.

LaDeau, S. L., A. M. Kilpatrick, and P. P. Marra. 2007. West Nile virus emergence and large-scale declines of North American bird populations. Nature 447:710-U713.

LaDeau, S. L., P. P. Marra, A. M. Kilpatrick, and C. A. Calder. 2008. West Nile Virus Revisited: Consequences for North American Ecology. BioScience 58:937-946.

Lardeux, F. J., R. H. Tejerina, V. Quispe, and T. K. Chavez. 2008. A physiological time analysis of the duration of the gonotrophic cycle of Anopheles pseudopunctipennis and its implications for malaria transmission in Bolivia. Malaria Journal 7.

Lounibos, L. P. 2002. Invasions by insect vectors of human disease. Annual Review of Entomology 47:233-266.

Marra, P. P., S. Griffing, C. Caffrey, A. M. Kilpatrick, R. McLean, C. Brand, E. M. I. Saito, A. P. Dupuis, L. Kramer, and R. Novak. 2004. West Nile Virus and Wildlife. BioScience 54:393-402.

Mordecai, E. A., K. P. Paaijmans, L. R. Johnson, C. Balzer, T. Ben-Horin, E. Moor, A. McNally, S. Pawar, S. J. Ryan, T. C. Smith, and K. D. Lafferty. 2013. Optimal temperature for malaria transmission is dramatically lower than previously predicted. Ecology Letters 16:22-30.

Morris, C. D. 1989. Eastern Equine encephalitis. Pages 1-20 in T. P. Monath, editor. The arboviruses: epidemiology and ecology. Boca Raton, Fla, CRC Press.

Murray, C. J. L., L. C. Rosenfeld, S. S. Lim, K. G. Andrews, K. J. Foreman, D. Haring, N. Fullman, M. Naghavi, R. Lozano, and A. D. Lopez. 2012. Global malaria mortality between 1980 and 2010: a systematic analysis. Lancet 379:413-431.

Paaijmans, K. P., S. Blanford, B. H. K. Chan, and M. B. Thomas. 2012. Warmer temperatures reduce the vectorial capacity of malaria mosquitoes. Biology Letters 8:465-468.

Patz, J. A., M. A. McGeehin, S. M. Bernard, K. L. Ebi, P. R. Epstein, A. Grambsch, D. J. Gubler, P. Reither, I. Romieu, and J. B. Rose. 2000. The potential health impacts of climate variability and change for the United States: Executive summary of the report of the health sector of the US national assessment. Environmental Health Perspectives 108:367.

Paweska, J. T. and P. J. van Vuren. 2014. Rift Valley Fever virus: A virus with potential for global emergence. Pages 169-200 in N. Johnson, editor. Role of Animals in Emerging Viral Diseases. Elsevier, London.

Reeves, W. C., J. L. Hardy, W. K. Reisen, and M. M. Milby. 1994. Potential effect of global warming on mosquito-borne arboviruses. Journal of Medical Entomology 31:323-332.

Reisen, W. K., Y. Fang, and V. M. Martinez. 2006. Effects of temperature on the transmission of West Nile virus by Culex tarsalis (Diptera : Culicidae). Journal of Medical Entomology 43:309-317.

Reisen, W. K. 2010. Landscape Epidemiology of Vector-Borne Diseases. Annual Review of Entomology 55:461-483.

Reiter, P. 2001. Climate change and mosquito-borne disease. Environmental Health Perspectives 109:141-161.

Scott, T. W. and S. C. Weaver. 1989. Eastern Equine encephalomyelitis virus: Epidemiology and evolution of mosquito transmission. Advances in Virus Research 37:277-328.

Shope, R. E. 1992. Impacts of global climate change on human health: Spread of infectious disease. Pages 363-370 in S. K. Majumdar, L. S. Kalkstein, B. M. Yarnal, E. W. Miller, and L. M. Rosenfeld, editors. Global Climate Change: Implications, Challenges and Mitigation Measures. Pennsylvania Academy of Science, Easton.

Sutherst, R. W. 2004. Global change and human vulnerability to vector-borne diseases. Clinical Microbiology Reviews 17:136-173.

Tatem, A. J., S. I. Hay, and D. J. Rogers. 2006. Global traffic and disease vector dispersal. Proceedings of the National Academy of Sciences of the United States of America 103:6242-6247.

van Riper, C., S. G. van Riper, M. L. Goff, and M. Laird. 1986. The epizootiology and ecological significance of malaria in Hawaiian land birds. Ecological Monographs 56:327-344.

Watts, D. M., D. S. Burke, B. A. Harrison, R. E. Whitmire, and A. Nisalak. 1987. Effect of temperature on the vector efficiency of Aedes aegypti for dengue-2 virus. American Journal of Tropical Medicine and Hygiene 36:143-152.

Weaver, S. C. and W. K. Reisen. 2010. Present and future arboviral threats. Antiviral research 85:328-345.