Paper SA-10-2015

Improving Data Collection Efficiency And Programming Process Flow Using SAS® Enterprise Guide® With Epi Info™

Chad Wetzel, MPH, Douglas County Health Department, Omaha, NE
Dr. Anne O'Keefe, MD, MPH, Douglas County Health Department, Omaha, NE
Justin Frederick, MPH, Douglas County Health Department, Omaha, NE

ABSTRACT

Epi Info™ is a commonly used software program used by public health professionals for data collection, analysis, and reporting. There are many benefits of using Epi Info in conjunction with SAS. Using each of these two systems together enables users to benefit from the strengths of both to improve the efficiency of managing and analyzing public health data. Epi Info provides an easy to use data entry form for manual data entry and SAS provides robust capabilities to manage, analyze, and report data. SAS® Enterprise Guide® simplifies the complexities needed for running multiple SAS programs. In order to combine public health data reported electronically with manually entered data, numerous lines of SAS code are required to match variables and formats that can be converted to and from Epi Info. It also requires a complex order of sorting and combining of data subsets in order to accurately and efficiently extract data from both systems, update data entries with all available data, and remove duplicated data entries. This paper focuses on how SAS Enterprise Guide better organizes the multiple steps required for complex projects, increases efficiency of data collection, creates intermediate data sets needing to be monitored, and generates epidemiological reports using output from multiple SAS programs.

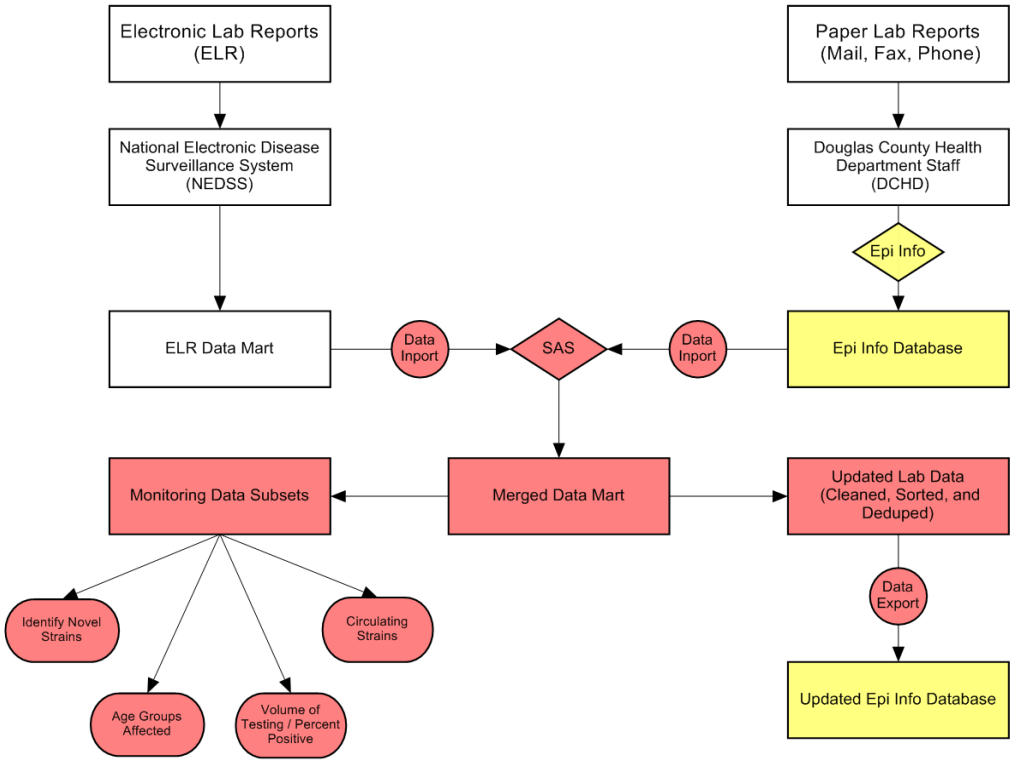
INTRODUCTION

Influenza is one of the most commonly reported communicable diseases. Influenza reports are sent to local health departments in a variety of ways. Health departments may receive anywhere from hundreds to thousands of influenza reports every year. During the 2014-15 Influenza Season, more than 25,000 electronic laboratory reports for influenza were received by Douglas County through the National Electronic Disease Surveillance System (NEDSS) and more than 1,900 laboratory reports were received on paper. This includes both positive and negative test results. Multiple methods of reporting by laboratories and healthcare providers creates many challenges for effective disease surveillance. Laboratories use multiple means of communication, send reports in different formats, and frequently send duplicate and/or updated reports. Duplicate and updated reports result from both multiple labs reporting the same result (due to laboratory send outs) and laboratories reporting the same result in multiple methods (electronically, by fax, and/or by mail).

Electronic reporting in computer information systems is becoming the predominate method of receiving laboratory reports. SAS is commonly used by epidemiologists to manage and analyze electronic laboratory reports. Laboratory reports sent by mail, fax, or called in over the phone can easily be collected using Epi Info due to the ease of data entry. Merging electronically reported laboratory reports with manually entered reports into one database helps to more efficiently manage and analyze all laboratory reports without requiring dual entries in both systems. It significantly reduces time needed for staff to enter data and provides for more accurate influenza surveillance activities.

This paper shows how SAS Enterprise Guide enables epidemiologists to merge laboratory data reported electronically with data reported on paper, which optimizes influenza surveillance activities by enhancing accuracy and increasing efficiency. It will also focus on how SAS Enterprise Guide improves the organization of multiple SAS programs using the process flow feature. This allows epidemiologists to easily and rapidly clean and monitor subsets of data useful for important influenza surveillance activities such as detecting novel strains of influenza, monitoring the frequency of testing, and identifying reporting errors or new methods of reporting.

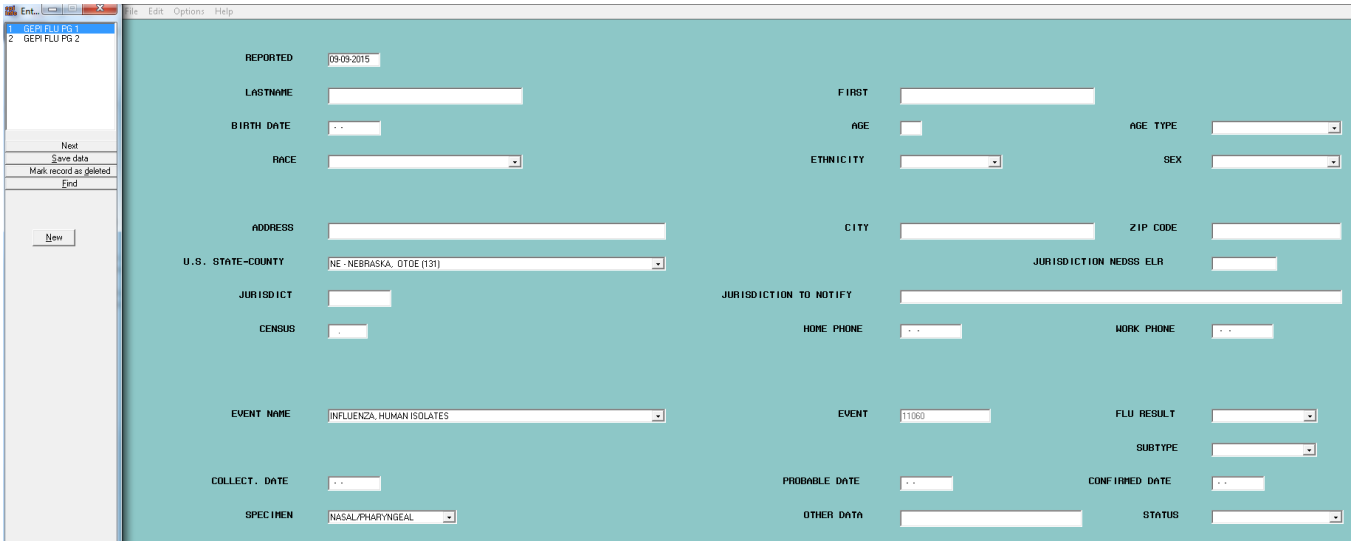
Figure 1: Flowchart for managing reported influenza tests



EPI INFO DATABASE

Epi Info (http://www.cdc.gov/epiinfo/) is a free software application developed by the Centers for Disease Control and Prevention (CDC) as a tool for epidemiologists to create questionnaires, enter, and analyze data. Epi Info provides the ability to design databases that enables public health professionals to easily enter data for laboratory reports. Figure 2 below shows a portion of the database designed to capture data from laboratory reports for influenza.

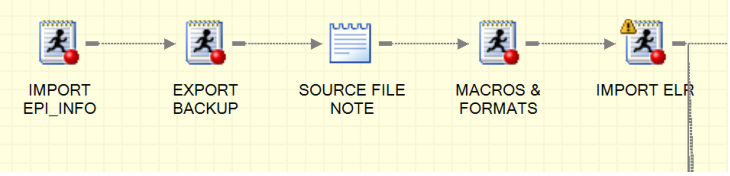
Figure 2: Epi Info influenza database – data entry form



SAVING BACKUP COPIES

Each time the influenza database is updated, a copy of the Microsoft® Access file from Epi Info is saved in a separate folder using the Export procedure in order to protect any corruptions or errors that may occur during the update process. The most recent data set of electronic laboratory reports is then imported and all-non influenza reports are removed using several variables describing the type of laboratory test that was conducted and the result of the test.

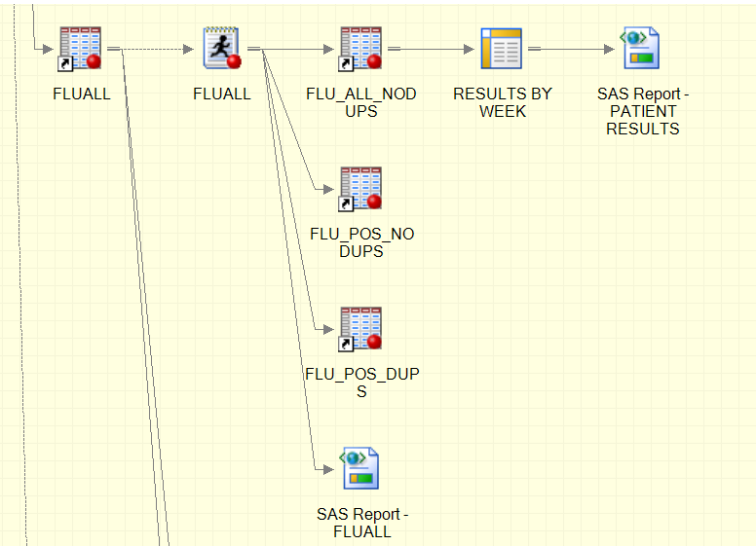
Figure 3: Process flow for importing new data and saving a backup copy of the database



MONITORING SUBSETS OF DATA

Several subsets of data are created in order to monitor influenza testing and methods of reporting. A subset that includes all individuals tested for influenza and those that tested positive is created by deduplication of the data set by name and birthdate using the Sort procedure. Using the Freq procedure enables epidemiologists to monitor the number of unique individuals being tested for influenza in the community and those that tested positive.

Figure 4: Process flow for monitoring the frequency of reported influenza tests




```
*Creating a subset of all individuals tested and #/% of positive tests;

      proc sort data=fluall out=flu_all_nodups nodupkey;
      by patient_last_name patient_first_name dob descending result;
      run;

      proc freq data= flu_all_nodups;
      tables result;
      run;
```

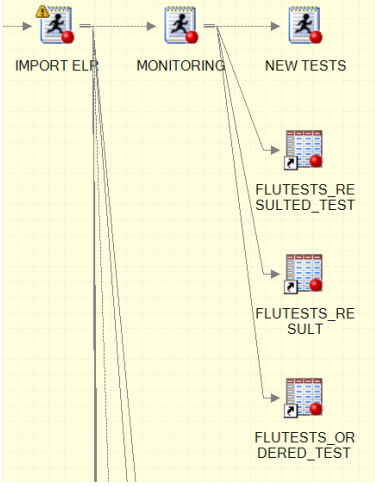
Another subset is created to monitor how laboratories are reporting influenza tests. All variables that would indicate an influenza test are monitored for new incoming results. This is conducted using the FIND function in the DATA step.

```
*Creating a subset of any tests with 'FLU' in test variables and removes
tests with text strings not associated with influenza testing;

      data flutests;
      set fluseason;
      tests  = find(result,'FLU');
      tests1 = find(resulted_test,'FLU');
      tests2 = find(result,'PARAINFLUENZA');
      tests3 = find(resulted_test,'PARAINFLUENZA');
      tests4 = find(result,'HAEM');
      tests5 = find(resulted_test,'HAEM');
      tests6 = find(result,'TITER');
      tests7 = find(resulted_test,'TITER');
      tests8 = find(result,'STREP');
      tests9 = find(result,'NEGATIVE FOR INFLUENZA');
      tests10 = find(result,'HEMOPHILUS');
      tests11 = find(result,'PSEUDO');
      tests12 = find(resulted_test,'AB.IGG');
      tests13 = find(resulted_test,'AB. IGG');
      tests14 = find(resulted_test,'ANTIBODY');
      run;

      data flutests1;
      set flutests;
      if tests > 0 or tests1 > 0;
      if tests2 > 0 then delete;
      ...
      if tests14 > 0 then delete;
      run;
```

All known data values are deleted from a subset of data for each monitored variable. If a new data value is identified in any of the monitored variables (resulted_test, result, ordered_test), PROC FREQ will identify any new data value reported in those variables. These data subsets also enables epidemiologists to identify if any novel influenza strains are reported by identifying new flu results not reported previously.

```
*Example: Deleting all known data values associated with influenza testing
for the resulted_test variable;

      data flutests2;
      set flutests1;
      if
      (resulted_test='INFLU A SCREEN RESULT') or
      ...
      (resulted_test='INFLUENZA B ANTIGEN') then delete;
      run;

      proc freq data=flutests2;
      tables resulted_test;
      run;
```

Figure 5: Process flow for monitoring data values for results of reported influenza tests



Other subsets are created to monitor lab reports with missing address information as well as potential data entry errors. The Print procedure is used to identify multiple lab reports with the same name and date of birth results.

```
*Example: Identifying observations with not enough address information to
determine county of residence;

      data elr_nodem1;
      set combined4;
      if statecountycode = 'NE - NEBRASKA,  UNK NEBRASKA COUNTY (999)';
      run;

*Example: Monitoring potential errors in entering the spelling of last name
in order to prevent the creation of 2 separate lab reports;

      proc freq data=combined4;
      tables first*birthdate / noprint out=first_birthdate;
      run;

      proc print data=first_birthdate;
      where count ge 2;
      run;
```

RECODING AND NEW VARIABLES

The data set that includes all electronic laboratory reports requires a series of steps to clean the data. Several variables need created or recoded to include pertinent information needed for influenza surveillance. Variables and data values must match those in the Epi Info database. This requires a lengthy set of IF-THEN-ELSE statements.

```
*Example: Create a new county variable using patient zip code;

      data flutype4;
      set flutype3;
      if patient_zip1 = 68001 then county = 'BUTLER';
      if patient_zip1 = 68002 then county = 'WASHINGTON';
      ...
      run;
```

The program also uses a series of text searches to determine what type of influenza test was conducted. The program searches several text strings (SUBTYPING, RNA, PCR, etc.) using the FIND statement to identify if it was a PCR test and if the test identified a specific subtype (H1, H3, H1N1, etc.).

FORMATTING AND MERGING DATA

Each variable must be renamed and formatted to match the variables in the Epi Info database. The data values also have to match that of Epi Info because many of the variables in Epi Info are drop down options and not open text fields. Also, any variables not in the Epi Info database are deleted. The updated data set is then merged with the Epi Info database.

```
address = patient_street_address;
age = ageyrs;
birthdate =dob;
...
if patient_current_sex = 'MALE' then sex = 'M';
if patient_current_sex = 'FEMALE' then sex = 'F';
```

CREATING AN INFLUENZA TEST HIERARCHY

One of the difficulties with laboratory reporting is that several different laboratory reports come in for the same person. Also, many times a person gets tested using more than one type of test. For example, if a rapid influenza test is received from a lab as well as a PCR test, the results of the PCR are more likely to be correct and override the results of a rapid influenza test. These same hierarchy variables are added to both electronic and manually entered laboratory reports. After the data is merged together, the laboratory report with the most accurate test type is kept in the database (The lower the number, the more accurate the test or diagnosis).

```
if flu = 'ANTIGEN A' then flu1 = '5A';
if flu = 'ANTIGEN B' then flu1 = '5B';
if flu = 'TESTPCR A' and subtype = ' ' then flu1 = '3A';
if flu = 'TESTPCR A' and subtype = 'H3' then flu1 = '2A';
if flu = 'TESTPCR A' and subtype = 'H3N2' then flu1 = '1A';
if flu = 'TESTPCR A' and subtype = 'H3N2d' then flu1 = '1A';
if flu = 'TESTPCR A' and subtype = 'H1N1' then flu1 = '1A';
if flu = 'TESTPCR A' and subtype = 'H1' then flu1 = '2A';
if flu = 'TESTPCR B' then flu1 = '3A';
if flu = 'CULTURE A' then flu1 = '4A';
if flu = 'CULTURE B' then flu1 = '4B';
if flu = 'ANTIGEN A/B' then flu1 = '1C';
if flu = 'OTHER' then flu1 = '1D';
if flu = 'CX DIAG' then flu1 = '1E';
```

SORTING AND DEDUPING

Once the data set is merged with the Epi Info database, the data set is sorted by name and date of birth, followed by the flu test hierarchy variable and home address. The data set is then deduped using PROC SORT in order to keep only the most accurate test result with the most specific home address information.

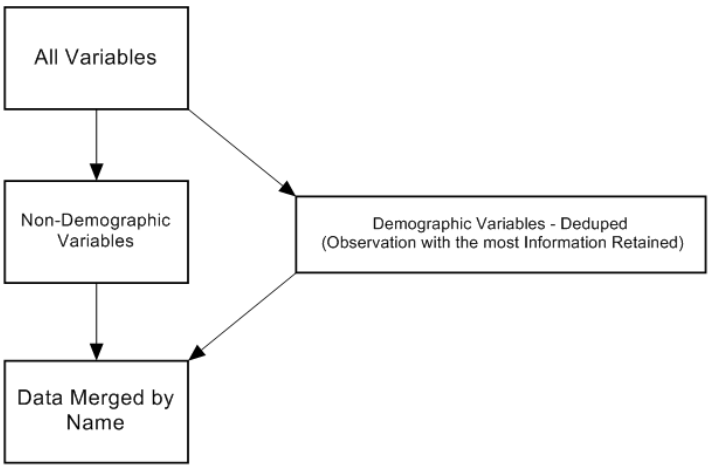
```
proc sort data=combined out=combined1;
by lastname first birthdate flu1 specimen1 descending address
descending city descending state;
run;

proc sort data=combined1 nodupkey out=combined2;
by lastname first birthdate;
format birthdate collected completed received reported mmddyy10.;
run;
```

EXTRACTING AND SAVING IMPORTANT DATA

Multiple lab reports on the same individual are commonly received electronically as well as through fax, mail, or by phone. It is not uncommon for certain data elements to be missing on paper but present in the electronic report or vice versa. The influenza program is designed to extract data from both manually entered lab reports as well as electronically received reports in order to obtain as much information as possible for surveillance. Demographic

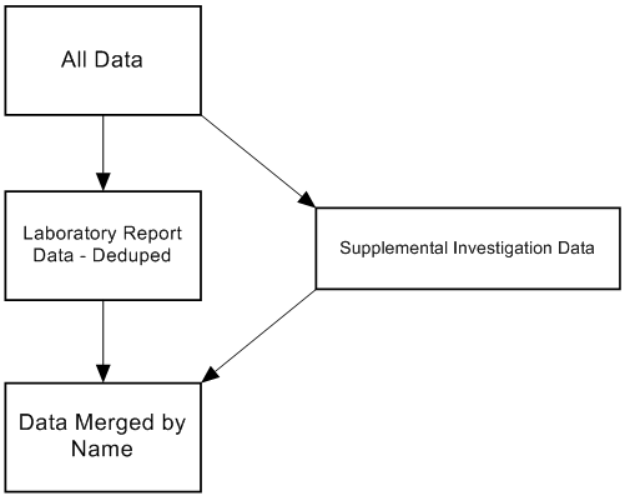
information such as address, phone number, and race/ethnicity status is important for public health surveillance. Separate sorting and deduplication procedures using PROC SORT for demographic information are necessary in order to extract as much data as possible. This is done separately for address/phone number and race/ethnicity because these are frequently not available in the same report or observation.

Figure 6: Flow diagram of how the observations with the most demographic information is extracted and saved by merging the most available information into all observations associated with a given name



Some influenza reports need follow-up investigations completed. This information is entered in a separate page in the Epi Info database. In order for this data to stay connected with the most up-to-date laboratory data on an individual, this set of variables must be separated during the sorting procedure and merged back once sorting and deduping is completed.

Figure 7: Flow diagram of how the observations with supplemental investigation data is extracted and saved by merging the investigation data into all observations associated with a given name.



EXPORTING DATA TO EPI INFO

Several steps must be completed before exporting the final data set into the Microsoft Access file created by Epi Info. Epi Info automatically creates a variable called “UniqueKey” that orders each observation in the database. This variable must be deleted prior to sorting the final data set. The data set is then sorted by report date before exporting. After the export process, a new “UniqueKey” is automatically generated in the order of report date

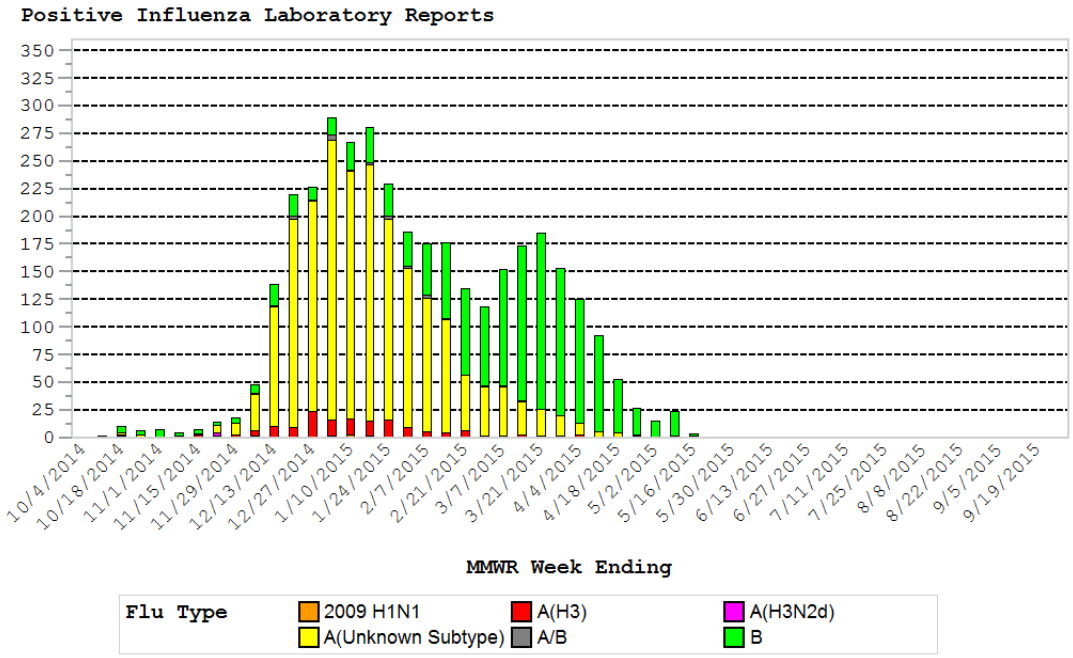
once the Epi Info database is reopened. A variable that must be retained in the data set before exporting is "RecStatus." This is an automatically generated variable created by Epi Info and must be retained in order for the observations to be included in the database. The final data set must also have date variables properly recoded to match the date format used in Epi Info. Several tables in a Microsoft Access file are created by Epi Info when a database is created. During the export process using PROC EXPORT, the name of the new copy of the table with the laboratory data must also match the name of the table in the Microsoft Access file created by Epi Info in order to override the previous copy. Once the export process is complete, the database is successfully updated. Due to the automatic creation of the UniqueKey" variable, the database may need to be opened, closed, and then reopened again before new "UniqueKeys" are generated by Epi Info.

```
proc sort data=combined4;
by reported;
run;

proc export data=combined4
      outtable= "FLU1516"
      dbms=accesscs replace;
      database=" H:\DC_EPI_INFO_FLU\FLU_DATABASE\FLU1516.MDB ";
run;
```

REPORTING

During the influenza season DCHD creates and disseminates a weekly influenza surveillance activity report to healthcare facilities, providers, and infection control practitioners with updates on influenza activity in the community. This report helps the healthcare community better prepare for and manage patients with influenza like illness. The table below is an example of one of the graphs that is created in SAS Enterprise Guide and included in the weekly influenza report. The most recent weekly influenza surveillance activity report can be found at: (http://www.douglascountyhealth.com/disease-a-immunization/influenza).

Figure 8: Example Graph – Number of Positive Influenza Laboratory Reports in Douglas County Residents by Flu Type, 2014-15 Season



CONCLUSION

During the 2014-2015 influenza season, more than 3,500 positive laboratory reports for influenza were received by the Douglas County Health Department. Of all laboratory reports received for influenza testing, 76% were reported electronically. Manually entering all laboratory reports for influenza is very resource intensive. Using SAS and Epi Info in conjunction has successfully reduced the amount of staff time required to maintain the influenza database for surveillance activities. It also has significantly improved the accuracy and completeness of reported data and enhanced the ability to monitor the frequency of influenza testing, emerging strains of influenza, and changes in laboratory reporting methods.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name: Chad Wetzel
Enterprise: Douglas County Health Department
Address: 1111 South 41st Street
City, State ZIP: Omaha, Nebraska 68105
Work Phone: 402-444-7214
Fax: 402-444-3287
E-mail: chad.wetzel@douglascounty-ne.gov
Web: www.douglascountyhealth.com