Paper EPO-188

Privacy Protection Using Base SAS®: Purging Sensitive Information from Free Text Emergency Room Data

Michelle White, Thomas Schroeder, Li Hui Chen, Jean Mah

U.S. Consumer Product Safety Commission

ABSTRACT

Federal agencies must balance privacy protection concerns with the competing priorities of accessibility and usability mandated by open data initiatives. Free text data often provide detail and qualitative value not offered by coded data. However, free text narratives are more likely to contain personally identifiable or sensitive information. This paper describes how U.S. Consumer Product Safety Commission (CPSC) staff strategically identifies sensitive information in hospital emergency department (ED) narratives using macros, Perl regular expressions and the PRXMATCH function in Base SAS® Version 9.

CPSC's National Electronic Injury Surveillance System (NEISS) is a national probability sample of hospitals with EDs in the United States and its territories. The NEISS collects information for about 400,000 product-related ED visits annually. Each NEISS record includes coded variables and a brief text narrative. This narrative may contain sensitive information (*e.g.,* patient names, product brands) that must be purged before the NEISS data are publicly released. About 65 percent of the narratives are immediately reviewed and purged, if necessary, by contract reviewers. All narratives are subsequently input into a SAS® program to identify potentially sensitive words in the remaining un-reviewed narratives and to evaluate the completeness of the review by contractors. A macro compares each narrative to a SAS® data set, where each observation contains a "purge term" and corresponding description. A "purge term" may be a literal string or Perl regular expression. Perl regular expressions are advantageous because they can encompass misspellings, keystroke errors, irregular spacing, or numerical identifiers in a specific format (*e.g.*, social security number, birthdate). If a "purge term" is contained in the narrative, then the case is output for review by CPSC staff.

Thus, CPSC staff reviews only narratives with a high probability of containing potentially sensitive information. Previously unreviewed narratives may be purged; and purges done on any narratives previously reviewed by contractors are marked as "missed purges" by the contractors. New "purge terms" are periodically identified by a SAS® program that compares the terms actually purged from reviewed narratives to those in the existing SAS® data set.

INTRODUCTION

NATIONAL ELECTRONIC INJURY SURVEILLANCE SYSTEM (NEISS)

For more than 30 years, the CPSC has operated a statistically valid injury surveillance system known as the National Electronic Injury Surveillance System (NEISS). The primary purpose of the NEISS is to collect timely data on consumer product-related injuries occurring in the United States and its territories. The NEISS is a sample of approximately 96 hospital EDs from which the CPSC receives injury reports on a daily basis. NEISS coders abstract information from relevant ED records and transmit the data on a flow basis to the CPSC over a secure Internet connection. The stratified, probability-based sample design of the NEISS also enables CPSC staff and other public health data users to compute national injury estimates by product type, age, sex and other variables.

Traditionally, the NEISS has collected about 400,000 consumer-product injury reports annually from its sample of EDs. In 2000, through an interagency agreement with the Centers for Disease Control and Prevention (CDC), the NEISS was expanded to collect all trauma injuries treated in EDs (regardless of product involvement) in a two-thirds subsample of its hospitals. As a result, the NEISS currently collects almost 800,000 injury reports annually to support the work of both agencies.

PUBLIC USE OF THE NEISS DATA

In addition to serving as the basis for statistical estimates, the NEISS data are also used by CPSC staff and public data users to provide information on how consumers interact with products leading up to an injury. Such information can be found in the 142-character free text narrative that is included in each NEISS case. Nearly every other variable on the NEISS record is a coded value (e.g., sex, diagnosis, body part, race, and disposition). Among its many uses, the narrative is used to perform quality control checks on the coded variables.

Beginning in 2003, the NEISS data became available for download from the CPSC's website and the need for redacting or purging the ED narratives of any personally identifiable or sensitive information became a major concern.

CPSC is considered a public health entity with access to medical records under the Health Insurance Portability and Accountability Act (HIPAA); however, all hospitals participate in the NEISS on a voluntary basis. Any inadvertent publishing of personally identifiable patient information in the publicly available NEISS data could result in a hospital curtailing its participation. Although personally identifiable information is often thought of as a name, it could also be an address, location (*e.g.*, a school), or a birthdate. Additionally, because brands or manufacturers associated with products are frequently mentioned in the ED narratives of interest to CPSC, that type of information must also be purged due to information disclosure requirements imposed by section 6 of the Consumer Product Safety Act (CPSA).

Along with the expansion of the NEISS in 2000 to collect all injuries in a two-thirds subsample of its hospital EDs, resources were also provided for contractors to review all narratives in the subsample for personally identifiable and sensitive information. If such information was found in the narrative by the contract reviewers, the narrative was manually purged of the non-releasable term(s) and a purged narrative was created. However, resources were not available to dedicate the same case-by-case review to the remaining NEISS cases. As a result, a process was created so that only those narratives with a high probability of containing information that should be purged would be output for review.

PURGING SENSITIVE INFORMATION FROM ED NARRATIVES

PROCESS OVERVIEW

An overview of the NEISS purge process is shown in Figure 1. About 65 percent of the NEISS record narratives are immediately reviewed and purged, if necessary, by contract reviewers. When a term is purged from the narrative, it is replaced by three asterisks ("***"). The purged narrative is saved separately and used for any public data release, while the original text of the narrative is kept for internal use. The initial dictionary of purge terms was built by comparing the original text of the narrative with the purged narrative.

The SAS® purging code created by CPSC staff and described in this paper, is used to verify the contractor review and to identify potentially sensitive words in the other 35 percent of unpurged NEISS record narratives. A macro using the PRXMATCH function compares each narrative to a SAS® data set containing the purge terms as Perl regular expressions. If a purge term is matched in the narrative, the record is output for manual review and possible purging by CPSC staff using Microsoft Access.

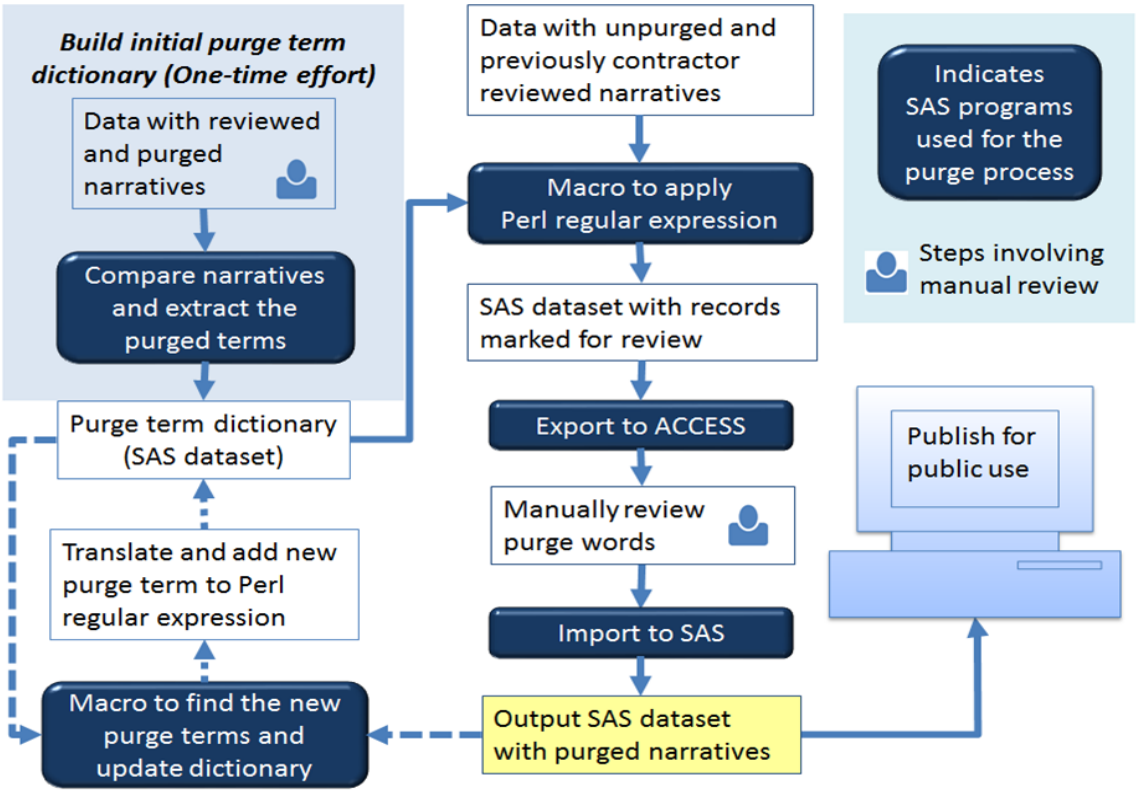

Figure 1. Overview of CPSC's NEISS Purge Process.

BUILDING THE DICTIONARY OF PURGE TERMS AND PERL REGULAR EXPRESSIONS

The dictionary of terms to be purged was originally created based on terms that were manually purged by reviewers; and the dictionary is periodically updated with new terms found by the contract reviewers and CPSC staff. This is done by extracting the purged terms from the original narrative using several SAS® character functions: COMBPL to remove double spaces; INDEX to find the starting position of the purged text; LENGTH to find the lengths of the original and purged narratives; and SUBSTR to extract the corresponding term from the original narrative. The method below assumes there is only one purged string in each narrative; however, the CALL PRXNEXT routine could be used to find multiple occurrences in one string.

```
DATA purgedterms (KEEP=narrative purgednarr purged_word);
   SET NEISSpurged;
/* Remove double-spaces */
   narrative = COMPBL(narrative);
   purgednarr = COMBPL(purgednarr);
   start = INDEX(purgednarr,'***');  /* Find location of purged string */
   IF start = 0 THEN DELETE;         /* Remove records without purged terms */
/* Identify the term that was purged */
   end_p = LENGTH(purgednarr);
   end_c = LENGTH(narrative);
   temp1 = SUBSTR(narrative,start,end_c - start +1);
   end_t = LENGTH(temp1);
   trim_p = end_p - start - 2;
   purged_word = SUBSTR(temp1,1,end_t - trim_p);
RUN;
```

narrative (Original Narrative)	purgednarr (Purged Narrative)	purged_word
UNK MALE HAD AN ACME ANVIL DROPPED ON HIS HEAD. DX: TBI	UNK MALE HAD AN *** ANVIL DROPPED ON HIS HEAD. DX: TBI	ACME
24YOM APPLYING SOUL-GLO TO HAIR AND DEVELOPED RASH	24YOM APPLYING *** TO HAIR AND DEVELOPED RASH	SOUL-GLO
43YOF CUT HER FINGER ON SOME PAPER FROM DUNDER MIFFLIN	43YOF CUT HER FINGER ON SOME PAPER FROM ***	DUNDER MIFFLIN

Table 1. Example of records from data set created by DATA step to extract purged terms

The dictionary is stored as a SAS® data set containing a description of the term and its associated Perl regular expression. A Perl regular expression is a string that describes a search pattern. Metacharacters (like wildcards but more versatile) can be used to find misspellings and typos. A single PERL regular expression can often replace what would require a long string of FIND or INDEX functions. A few examples of Perl regular expressions with metacharacters are given in Table 2 below. Several excellent resources on writing regular expressions and using them with SAS PRX functions are included in the References (Borowiak & Kenneth, 2007; Cody, 2003; Cody, 2010; Pless, 2004; Windham, 2014).

Term to be purged:	Regular expression:	Metacharacters:	Examples of Matches:
ACME	\WACME\W	\W matches any non-word/non-alphanumeric character excluding the underscore	“ACME” but not REPLACMENT
FLUBBER	FLUB+?ER	+? matches the ‘B’ one or more times)	FLUBER, FLUBBER
SOUL GLO	SOUL\W*?GLO	*? matches the \W zero or more times)	SOULGLO, SOUL-GLO, SOUL – GLO
DUNDER MIFFLIN	DUNDER\W*?M(F|I)+?LIN	(F|I)+? matches the F or the I one or more times	DUNDER-MIFILIN, DUNDERMFLIN
BIRTHDATE?	\d+?\/\d+?\/\d+	\d matches any digit 0-9, and \/ matches the backslash	1/1/2016, 1/1/16, 01/01/16

Table 2. Example of terms to be purged and their associated regular expressions

MACRO TO FIND PURGE TERMS IN THE NARRATIVES

CPSC's SAS® program, used to identify narratives with purge terms, includes several steps. First, a DATA _NULL_ step counts the total number of Perl regular expressions in the dictionary. Then PROC SQL reads the Perl regular expressions into macro variables.

```
/* Count the number of Perl regular expressions in the dictionary */
DATA _NULL_;
    IF 0 THEN SET purgewords NOBS=nobs;
    CALL SYMPUTX("numnames",nobs);
    STOP;
RUN;
%PUT TOTAL: &numnames;

/* Read Perl regular expressions and descriptions into macro variables */
PROC SQL NOPRINT;
    SELECT regex, description
           INTO   :word1 - :word&numnames,
                  :desc1 - :desc&numnames
    FROM purgewords;
QUIT;
```

Next, the PURGEFIND macro writes the code that searches the narratives in the NEISS records for the PERL regular expressions from the dictionary. The PRXMATCH function returns the first position in the narrative where the PERL regular expression is matched. If there is no match, the PRXMATCH function returns a zero. Each narrative with a matching purge term is output to a temporary dataset, along with the description of the term. The %SUPERQ function prevents the macro processor from attempting to resolve the Perl regular expression (*e.g.*, if a company name includes the ampersand).

```
/* Create code that searches for purge words */
%MACRO PURGEFIND;
    DATA temp;
           SET NEISS;
           FORMAT name $30.;
           %DO K=1 %TO &NUMNAMES;
           IF PRXMATCH("/%SUPERQ(word&k)/",narrative) THEN DO;
                        name="%SUPERQ(DESC&k)";
                        OUTPUT;
                 END;
           %END;
    RUN;
%MEND PURGEFIND;

%PURGEFIND;
```

The macro checks every Perl regular expression against every narrative. This results in some narratives with multiple records in the output data set. However, some of the matches will be false positives. Therefore, the next DATA step identifies and removes the more common "false positive" matches before a PROC SORT step is applied to remove duplicate NEISS records. The entire narrative is reviewed and manually purged by CPSC staff; thus, it does not matter which matched record is kept.

```
DATA temp1;
    SET temp;
    IF name = 'AUSTIN' and PRXMATCH('/EXHAUSTIN/',narrative) THEN DELETE;
    IF name = 'BOBBY' and PRXMATCH('/BOBBY\W?PIN/',narrative) THEN DELETE;
    IF name = 'CHARLIE' and PRXMATCH('/CHARLIE\W?HORSE/',narrative) THEN DELETE;
    IF name = 'CHUCK' and PRXMATCH('/(NUM|WOOD)CHUCK/',narrative) THEN DELETE;
    etc...
RUN;

PROC SORT data=temp1 NODUPKEY;
    BY uniqueID;
RUN;
```

uniqueID	narrative	Name
160000001	UNK MALE HAD AN ACME ANVIL DROPPED ON HIS HEAD. DX: TBI	ACME
160000004	24YOM APPLYING SOUL-GLO TO HAIR AND DEVELOPED RASH	SOUL-GLO
160000010	43YOF CUT HER FINGER ON SOME PAPER FROM DUNDER MIFFLIN	DUNDER MIFFLIN
160000036	33YOM SEEN 10/02/2016 FOR ANKLE SPRAIN, FELL OFF SKATEBOARD TODAY, NOW LEG SWELLING. DX: TIBIA FRACTURE	BIRTHDATE?

Table 3. Example of records from TEMP1 data set

EXPORT TO ACCESS FOR MANUAL REVIEW AND PURGING

The records are then exported to Microsoft Access using PROC EXPORT. A CPSC staff member then reviews and manually purges each narrative using a form in Microsoft Access (Figure 2 below). The reviewer marks whether the narrative requires purging ('Purge?'). If purging is required, the reviewer replaces the term to be removed with three asterisks.

```
PROC EXPORT
    DATA=WORK.TEMP1
    OUTTABLE="REVIEW"
    DBMS=ACCESS REPLACE;
    DATABASE="C:\DIRECTORY\PURGEREVIEW.ACCDB";
RUN;
```

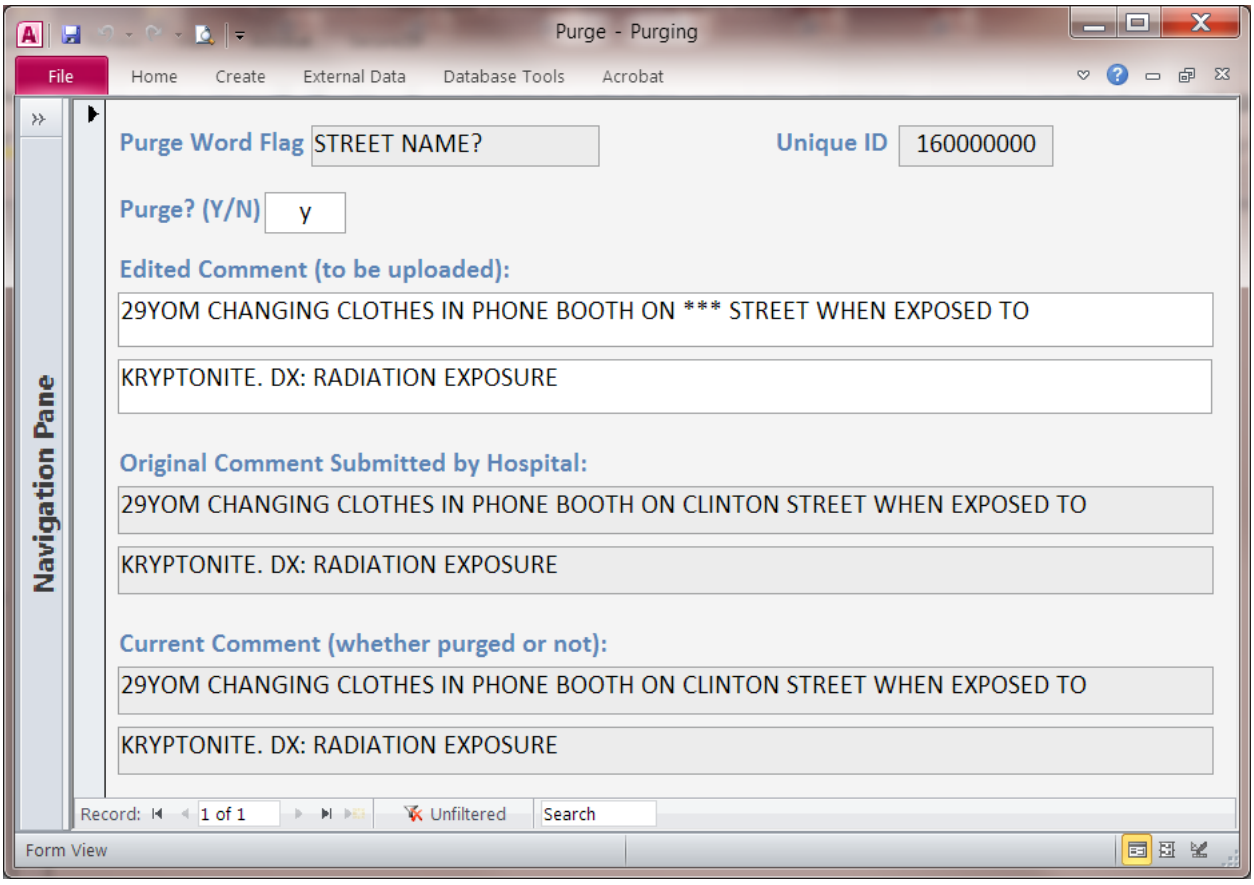

Figure 2. Microsoft Access Form for CPSC Staff to review and purge narratives

UPLOAD PURGED NARRATIVE AND PUBLISH FOR PUBLIC USE

The records are subsequently imported back into SAS® using PROC IMPORT; thereafter, the purged narratives are uploaded to the NEISS database. The purged narratives are now available for use in any publicly released data (Figure 3).

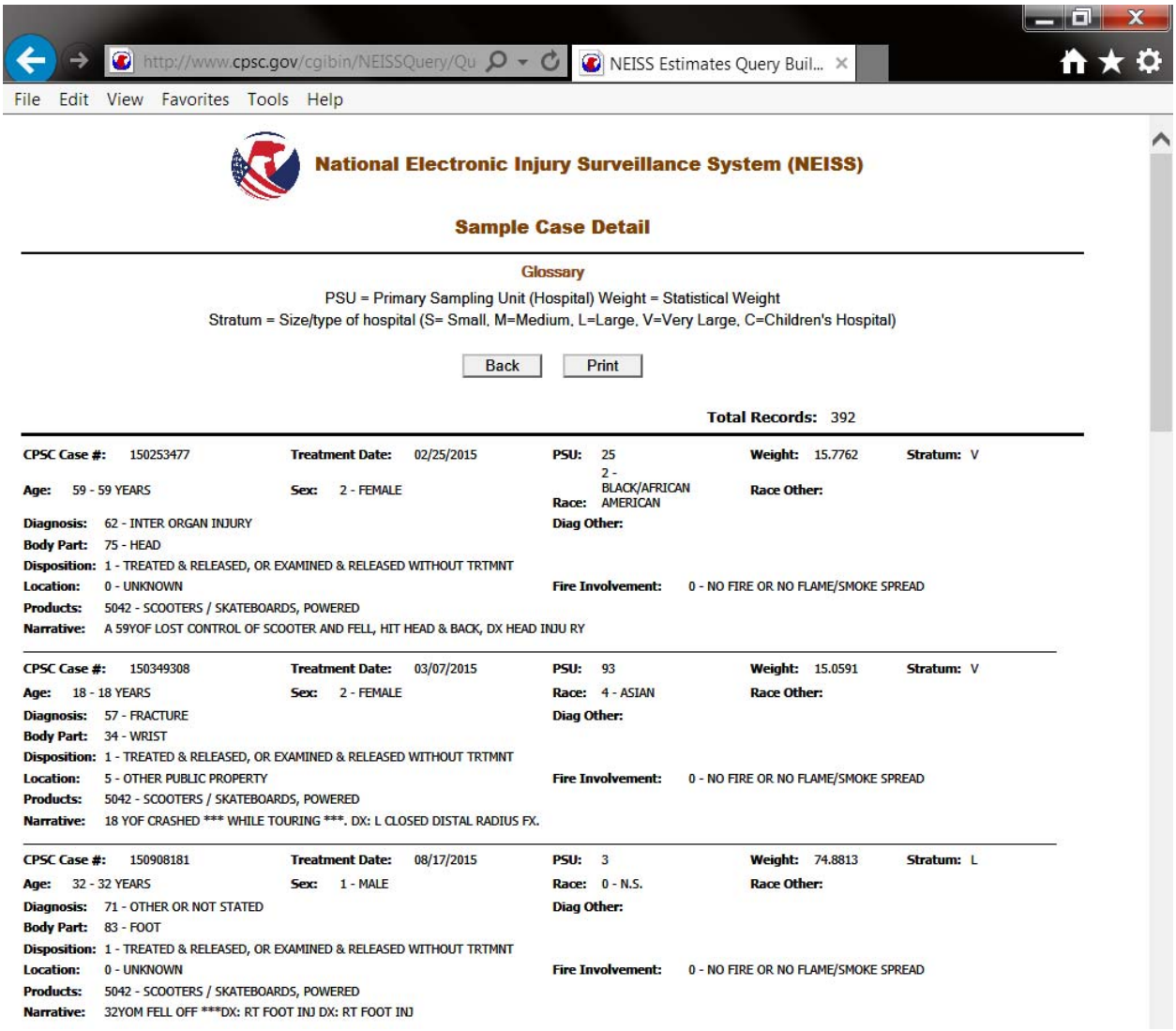

Figure 3. Screenshot of publicly available NEISS data with purged narratives, available at: http://www.cpsc.gov/cgibin/NEISSQuery/home.aspx.

TIPS AND TRICKS

- Remove double-spaces and trim leading and trailing blanks from your free text using one of the SAS® character functions (*e.g.*, COMBPL, CATS, TRIM), or account for their possibility in your PERL regular expressions (*e.g.*, include `\A\S*?` for possible leading blanks or `\S*?\Z` for possible trailing blanks.
- Perl regular expressions are not easy to read--always include a brief comment (or a separate variable) describing the Perl regular expression in plain language. Your future self and co-workers will thank you.
- When creating a dictionary of Perl regular expressions, consider how you will use the matched string once it is found. Perl regular expressions can be lazy (matching the shortest possible string) or greedy (matching the longest possible string). If you want to automatically replace the matched string, you may want to use a greedy search to match the entire string. For example, `\d+?\/\d+?\/\d+` will match 01/01/2000, but `\d+?\/\d+?\/\d+?` will stop at 01/01/2, and the remaining 000 will not be included in the matched string.

- The NEISS narratives are all uppercase. If your free text includes a mix of upper and lower case, you can use the 'i' option after the last delimiter in the PRX function to ignore case. For example, `prxmatch("/Widgets/i",narrative)` will match Widgets, WIDGETS, widgets, and WIdgets.

CONCLUSION

Free text data often provide detail and qualitative value that is not offered by coded data. Particularly for the CPSC and other public health data users, the NEISS narrative data are more useful to identify the injury scenarios in which patients interact with products. However, the usefulness of free text data must be balanced by the privacy protection concerns of patients and manufacturers. Therefore, it is incumbent on federal agencies to put systems in place to guard against the public release of private and sensitive information in publicly available data sets.

The manual review of all data records is not usually feasible or foolproof. Conversely, a process that solely depends on automation is not yet feasible, due to the huge variation of language and keystroking in the narratives. Therefore, a process that combines manual review with the assistance of computer programs has been developed at CPSC to meet the purging needs of the NEISS data.

Using tools available in Base SAS®, CPSC staff has semi-automated the purging of sensitive information from ED narratives. A similar process could be applied to other data containing free text, such as electronic medical records and death certificates.

REFERENCES

Borowiak, Kenneth. 2007. "Perl Regular Expressions 102." *Proceedings of the SAS Global Forum 2007*. Available at: http://www2.sas.com/proceedings/forum2007/135-2007.pdf.

Cody, Ron. 2004. "An Introduction to Perl Regular Expressions in SAS 9®." *Proceedings of the Twenty-Ninth SAS Users Group International Conference*. Available at: http://www2.sas.com/proceedings/sugi29/265-29.pdf.

Cody, Ron. 2010. *SAS® Functions by Example, Second Edition*. 445 pages. Cary, NC: SAS Institute Inc. Available at: https://www.sas.com/store/prodBK_62857_en.html.

Pless, Richard. 2004. "An Introduction to Perl Regular Expressions with Examples from Clinical Data." *Proceedings of the Twenty-Ninth SAS Users Group International Conference*. Available at: http://www2.sas.com/proceedings/sugi29/265-29.pdf.

Windham, Matthew. 2014. *Introduction to Regular Expressions in SAS®.* 120 pages. Cary, NC: SAS Institute Inc. Available at: https://www.sas.com/store/books/categories/usage-and-reference/introduction-to-regular-expressions-in-sas-/prodBK_67098_en.html.

RECOMMENDED READING

- A full list of metacharacters available for use with Perl regular expressions available in SAS® 9.4 is online at: http://support.sas.com/documentation/cdl/en/lefunctionsref/67960/HTML/default/viewer.htm#titlepage.htm.
- Information about NEISS is available at: http://www.cpsc.gov/en/Research--Statistics/NEISS-Injury-Data/.

CONTACT INFORMATION

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

Michelle White
U.S. Consumer Product Safety Commission
4330 East West Highway
Bethesda, MD 20814
mjwhite@cpsc.gov

DISCLAIMER

The views expressed in this paper are those of the authors and do not represent the views of the Commission.