

**Software User Manual**

**ETHER: Event-based Text-mining of Health Electronic Records**

***Submitted to***

**U. S. Food and Drug Administration**

**Center for Biologics Evaluation and Research**

***Submitted by***

**Engility Corporation**

[www.engilitycorp.com](http://www.engilitycorp.com)

**Sep. 15th, 2014**

Contents

[Table of Figures 3](#_Toc400453649)

[1. Introduction 4](#_Toc400453650)

[2. Install and Run ETHER software 4](#_Toc400453651)

[3. Getting Started 5](#_Toc400453652)

[3.1. Description of Extracted Features 5](#_Toc400453653)

[3.2. The Report Header 5](#_Toc400453654)

[4. ETHER Menu 6](#_Toc400453655)

[4.1. File Menu 6](#_Toc400453656)

[4.2. Data Menu 7](#_Toc400453657)

[4.3. Options Menu 7](#_Toc400453658)

[4.4. Help Menu 7](#_Toc400453659)

[5. Summarized Cases 8](#_Toc400453660)

[6. Case Narrative & Features 9](#_Toc400453661)

[7. Case Features over Time 9](#_Toc400453662)

[7.1 Plot Type 9](#_Toc400453663)

[7.2 Feature Selection 11](#_Toc400453664)

[7.3 Time Axis 11](#_Toc400453665)

[7.4 Time Range 11](#_Toc400453666)

[7.5 Auxiliary Tools 11](#_Toc400453667)

[7.5.1 Tooltip 11](#_Toc400453668)

[7.5.2 Text Highlight 11](#_Toc400453669)

[8 Group Plot over Time 12](#_Toc400453670)

[8.1 Lasagna Plot 12](#_Toc400453671)

[8.2 Feature Count Plot 12](#_Toc400453672)

[9 Data Filtering 13](#_Toc400453673)

[9.1 SMQ Tool 13](#_Toc400453674)

[9.2 Filter Builder 14](#_Toc400453675)

[10 System Preferences 14](#_Toc400453676)

# Table of Figures

[Figure 1: File Menu 6](file:///C:\CBERSciTool\TextMiner\doc\ETHER%20manual.docx#_Toc400453677)

[Figure 2: Data Menu 7](file:///C:\CBERSciTool\TextMiner\doc\ETHER%20manual.docx#_Toc400453678)

[Figure 3: Options Menu 7](file:///C:\CBERSciTool\TextMiner\doc\ETHER%20manual.docx#_Toc400453679)

[Figure 4: Summarized Cases Tab 8](#_Toc400453680)

[Figure 5: Case Narrative and Features Tab 9](#_Toc400453681)

[Figure 6: Visualization of features in text over time 10](#_Toc400453682)

[Figure 7: Visualization of features in split text over time 10](#_Toc400453683)

[Figure 8: Line plot for the number of features 11](#_Toc400453684)

[Figure 9: Lasagna plot 12](#_Toc400453685)

[Figure 10: Feature Count Plot 13](#_Toc400453686)

[Figure 11: SMQ filter window 14](#_Toc400453687)

[Figure 12: Filter builder 14](#_Toc400453688)

[Figure 13: System Preferences Dialog 15](#_Toc400453689)

# Introduction

Adverse event reports relating to biologics are made through several different adverse event reporting systems: the Vaccine Adverse Event Reporting System (VAERS), and the FDA Adverse Event Reporting System (FAERS). Other sources of unstructured free text information with surveillance interest are the blood mortality reports, the pre-clinical adverse event reports and the AIMS database with the public comments.

In the past, all adverse event reports related to biologics were directed to epidemiologists who reviewed them and sorted the reports. The sorting was based on the likelihood of any relationship of adverse events and use of the biologic that was implicated. Adverse events that were most likely to be related to the biologic were investigated further. The process of reviewing the report was time consuming and inefficient as much of the report is unstructured text and most reports are not directly related to the use of the biologic. We are starting to use a new suite of tools that assists epidemiologists in prioritizing their review.

OBE seeks to enhance the text mining and network analysis surveillance methodologies improve the prospective surveillance architecture consisting of pre-programmed queries and reports that will standardize information extraction and allow for efficiencies in the review of incoming adverse event reports from FAERS and VAERS. OBE is also looking into the processing of the above untapped sources of free text that may minimize the effort spent by other experts and reviewers within CBER. The overall goal is the creation of a Decision Support Environment that will incorporate advanced methodologies for the processing of the available OBE (and potentially other) data, and support the retrieval of specific information to facilitate the medical experts' review.

The first release of the in-house system, namely VaeTM (Vaccine adverse event Text Mining System) adequately extracted key features from VAERS reports as this has been shown in a medical evaluation study that was performed by the medical experts at OBE. As already described in the literature (<http://www.ncbi.nlm.nih.gov/pubmed/23650490>), VaeTM extracts key information from reports, such as diagnosis, cause of death, and time to onset of the adverse event.

The second release of the system, namely ETHER (Event-based Text-mining of Health Electronic Records), also captures the temporal information related to the already extracted features, and it visualizes the output in certain plots, translate it into MedDRA terms and supports the creation of report networks based on that.

# Install and Run ETHER software

To install ETHER, open Windows Explorer, type in [\\Cbsfs03\OBE Central\](file:///\\Cbsfs03\OBE%20Central\VAERS_Text_Mining\%20VAERS_Text_Mining.zip), find the folder “VAERS\_Text\_Mining” folder and copy it (the entire folder with the same name) to your C drive. If that folder is already on your drive, you need to replace it with this new one.

To run ETHER, look for the ether.exe and double click to run it. The ETHER software reads data file “reports.txt” as input. There is a sample “reports.txt” file in the folder, which contains all the Gardasil cases that came with the “Multiple Sclerosis” PT.

If you want to process some other set of reports, run a VAERS ad hoc query, select “Run Query” and then in the new window click on the “Start Text Miner” button. Depending on the number of reports it may take a while to view the next popup window for Citrix Environment (please see below). Hit Cancel in that window and you can find that a new “reports.txt” file is generated.

The final version of ETHER will obviously work in a fully automated fashion and you will not have to do anything else but run the ad hoc query. There are some technical issues we are trying to resolve with OIM right now and this is the reason for the lack of automation.

**Hardware/Software System Requirements to Run AeNA**

* Windows operating system (tested on Windows 7 64-bit OS)
* RAM: 4.0GB. The memory size depends on the network data to be processed

# Getting Started

While launching, the ETHER will attempt to connect to a local database on the user’s computer. If the report to be loaded can be found in the database, the previously stored text mining results will be retrieved and displayed; otherwise, ETHER will proceed by extracting features from unstructured narrative.

The local database is written to the file *C:\VAERS\_Text\_Mining/vaerslocal.db* in SQLite format, and its content can be accessed with any SQLite tool or software library, e.g. SQLite Browser.

The ETHER software displays reports, basic metadata, and extracted features in five tabs: Summarized Cases, Case Narrative & Features, Case Features over Time, Lab Data, and MedDRA Terms.

# Description of Extracted Features

The ETHER software extracts eleven features, as shown in Table 1, from the unstructured text contained within VAERS reports using the controlled vocabulary and grammar rules.

The user has the ability to export either limited features or all features in CSV format from *File* menu of ETHER.

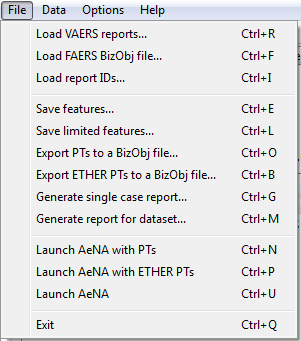
# The Report Header

The report header appears under the tab heading of every tab. It contains the following structured metadata retrieved from the VAERS database: Report ID, Age, Date of Exposure, Date of Onset, Vaccine Names, and MedDRA Preferred Terms.

Table 1: Medical Features Extracted by ETHER

|  |
| --- |
| **Feature** |
| CAUSE\_OF\_DEATH |
| DIAGNOSIS |
| DRUG |
| FAMILY\_HISTORY |
| MEDICAL\_HISTORY |
| TIME\_TO\_ONSET |
| LOT |
| RULE\_OUT |
| SECOND\_LEVEL\_DIAGNOSIS |
| SYMPTOM |
| VACCINE |

The user can change the currently displayed report by using the “Previous Report” and “Next Report” buttons.



# ETHER Menu

The main menu consists of four sections as follows:

# File Menu

Figure 1: File Menu

* *Load VAERS Reports*

Load report file generated by Vaers Explorer (VE).

* *Load FAERS BizObj file*

Load FAERS data in a Business Object file

* *Load report IDs*

Load reports to ETHER by a txt file with report IDs only.

* *Save features*

Save all features to a .csv file.

* Save limited features

Save a summary of features to a .csv file.

* *Export PTs to a BizObj file*

Save vaccines and PTs to a Business Object file which can be loaded in AeNA.

* *Export ETHER PTs to a BizObj file*

Save vaccines and PTs that are extracted from the narratives to a Business Object file, which can be loaded in AeNA.

* *Generate single case report*

Generate analysis report for single case

* *Generate report for dataset*

Generate analysis report for all cases in the dataset

* *Launch AeNA with PTs*

Launch AeNA with vaccines and PTs. The user needs to install the latest version of AeNA to launch it.

* *Launch AeNA with ETHER PTs*

Launch AeNA with vaccines and ETHER PTs.

* *Launch AENA*

Launch AeNA only.

* *Exit*

# Data Menu

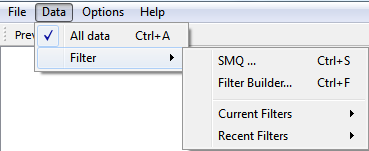
****

Figure 2: Data Menu

* **All Data**

Remove filter and restore the whole data set

* **Filter**
  + **SMQ**

Filter reports with SMQs (Standardised MedDRA query)

* + **Filter Builder**

Open filter builder

* + **Current Filters**

The filters being applied on the current data set

* + **Recent Filters**

Filters that have been applied recently. The filter can be reapplied.

# Options Menu

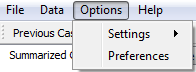


Figure 3: Options Menu

* **Settings**

*Color scheme*

Select color scheme for group plot

* **Preferences**

Open up the preference dialog

# Help Menu

Help document.

# Summarized Cases

The *Summarized Cases* Tab shows the summary of the current reports set. It provides an overview of reports. Each row represents a report with information of Case ID, Age, Gender, Vaccine, Calculated Onset, Alternative Explanation and Diagnosis.

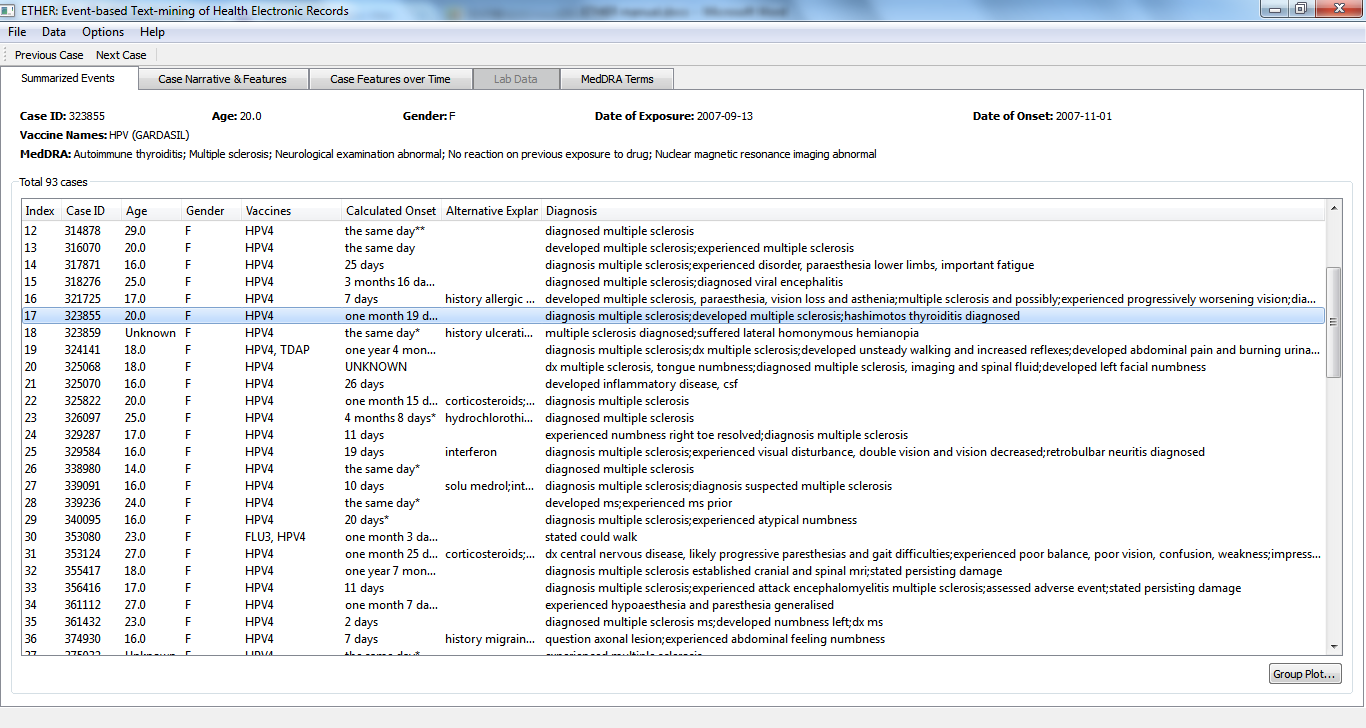


Figure 4: Summarized Cases Tab

Case ID, Age, Gender, and Vaccines are retrieved from report’s structured metadata in VAERS database, while the rest are extracted from the unstructured text of the report. The Diagnosis field represents concatenation of all PRIMARY\_DIAGNOSIS and CAUSE\_OF\_DEATH features extracted from the text of the report. If none exist, then all SECOND\_LEVEL\_ DIAGNOSIS features are shown instead. The Time to Onset column corresponds to TIME\_TO\_ONSET features extracted from the text. Finally, DRUG, MEDICAL\_HISTORY, and FAMILY\_HISTORY together features form Alternative Explanation column. The columns can be resized and sorted using the mouse. Double-clicking on the report will take the user to the Edit tab, on which the detailed view of the report is shown.

The onset time is calculated as follows:

1. When exposure and onset dates are provided from structured fields, use the provided dates, i.e., the input onset date minus exposure date
2. If dates are not provided, extract exposure and onset dates from text, denoted by \*;
3. If both 1) and (2) are not available, estimate the dates from text, denoted by \*\*.

The confidence level of the onset time is generally in the order of 1) > 2) > 3).

The reports can be sorted by clicking the table header in each column.

# Case Narrative & Features

This tab provides a detailed view of extracted features and narrative (VAERS report box 7) of a single case. The table shows feature type, feature text and time. The narrative is shown in the text box at the bottom. When a feature is selected, the corresponding text is highlighted in the text box.

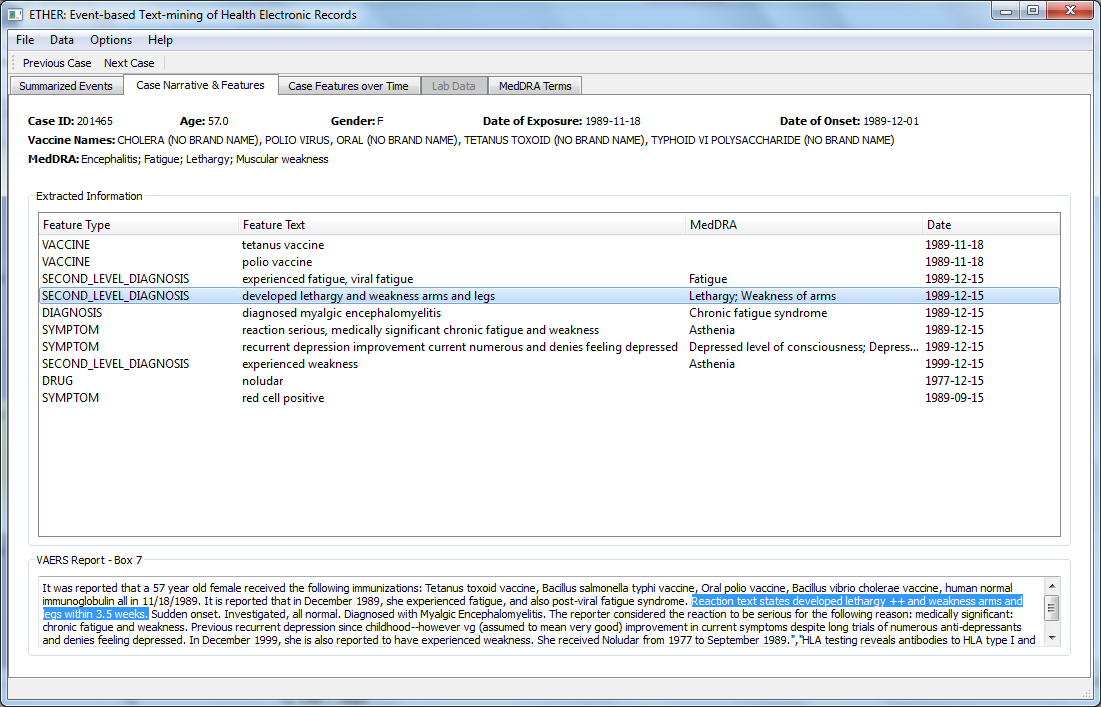


Figure 5: Case Narrative and Features Tab

# Case Features over Time

This tab provides a visualization tool for extracted medical features over time.

# Plot Type

There are three types for feature vs. time plots:

* Medical Feature Text

Feature text is shown in text boxes. Feature text may be trunked if it is too long to fit in the text box. User can hoop mouse over the feature text box to show the full feature text.

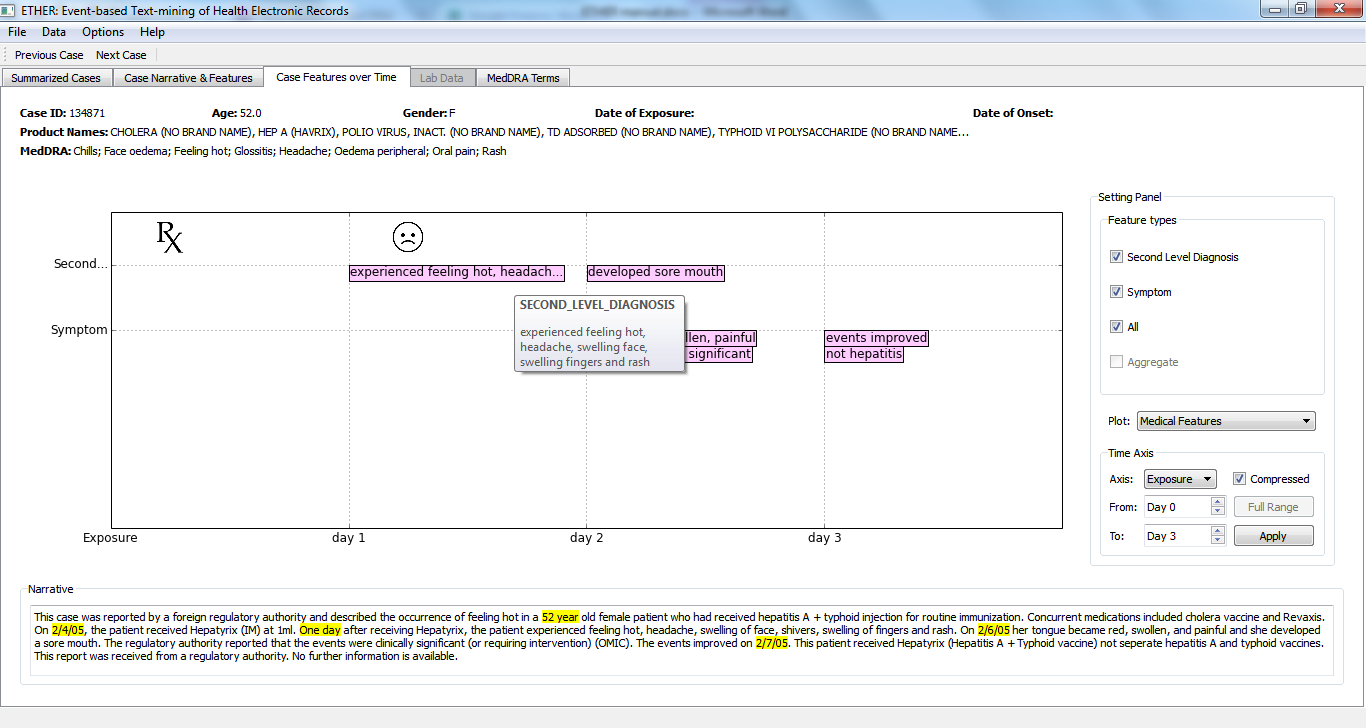


Figure 6: Visualization of features in text over time

* Split Feature Text

Composite feature text is split into multiple text boxes

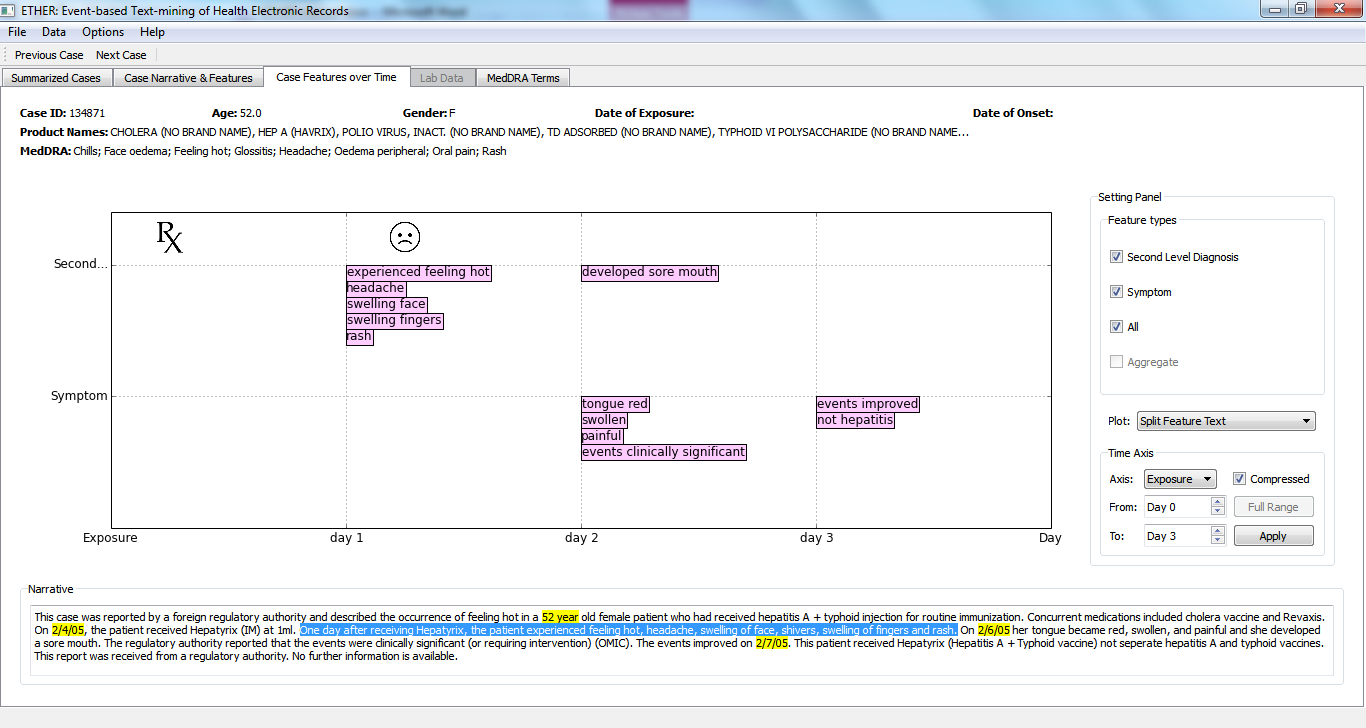


Figure 7: Visualization of features in split text over time

* Line Plot

Line plot shows the number of the features over time. Tooltip

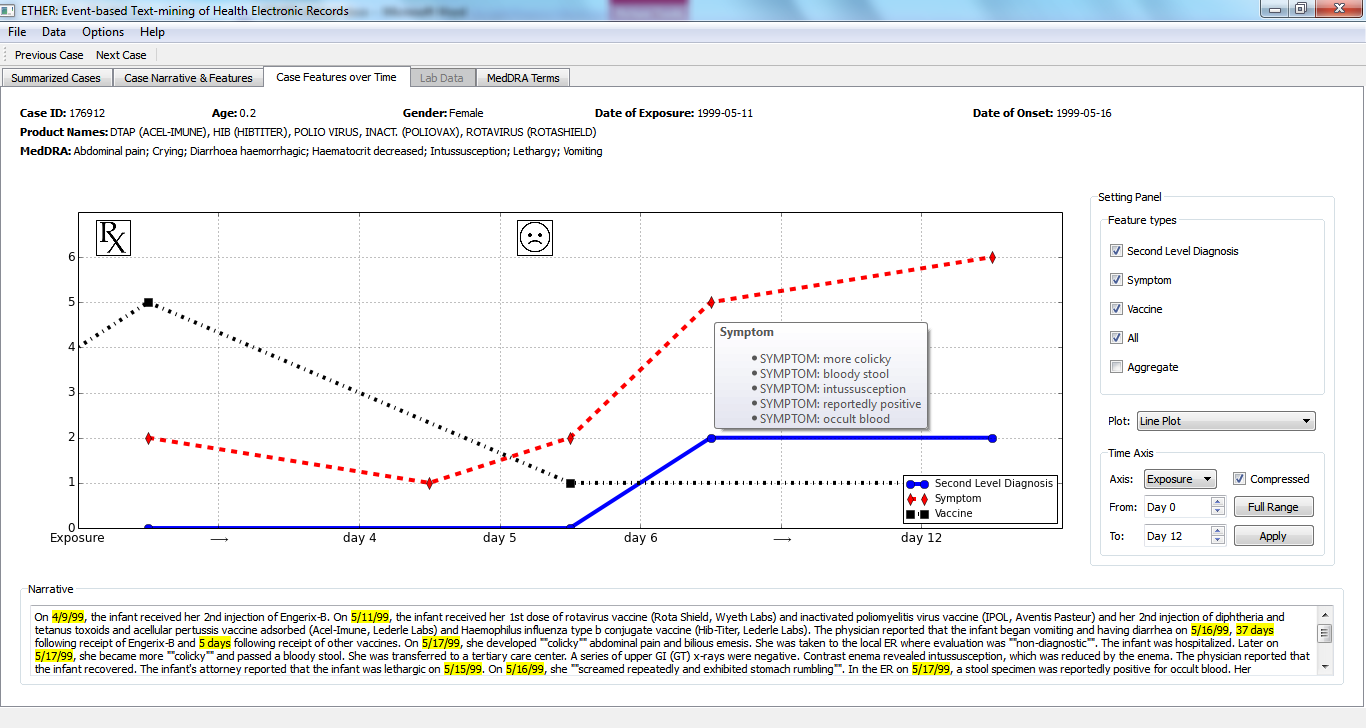


Figure 8: Line plot for the number of features

# Feature Selection

The user can select features to plot by check or uncheck the feature list in the setting panel.

# Time Axis

The user can select the time axis to be based on exposure date or calendar dates.

By default, the axis is compressed by removing large empty time intervals to use space more efficiently. By unchecking the compressed check box, the time axis is shown in its original scale.

# Time Range

The user can specify a time range in *From* and *To* boxes. When time axis is *Exposure*, the *From* and *To* boxes are spin box; when time axis is date, the edit boxes are date boxes.

When the current time range is full range, the “Full Range” button will be deactivated.

# Auxiliary Tools

# Tooltip

When mouse is hovering over a feature text or line plot nodes, detailed description is shown as a tooltip.

# Text Highlight

When a feature text box is clicked, its corresponding sentence is highlighted in the text box at the bottom.

# Group Plot over Time

Group plot for the current report set can be opened by clicking the *Group Plot* button in the *Summarized Cases* tab.

# Lasagna Plot

The Lasagna shows the number of medical features by colors, which can be quantified by the color bar on the right. In Figure 9, each row represents a single case over time. When mouse is hovering over a colored block, a list of features of this case is shown as a tooltip.

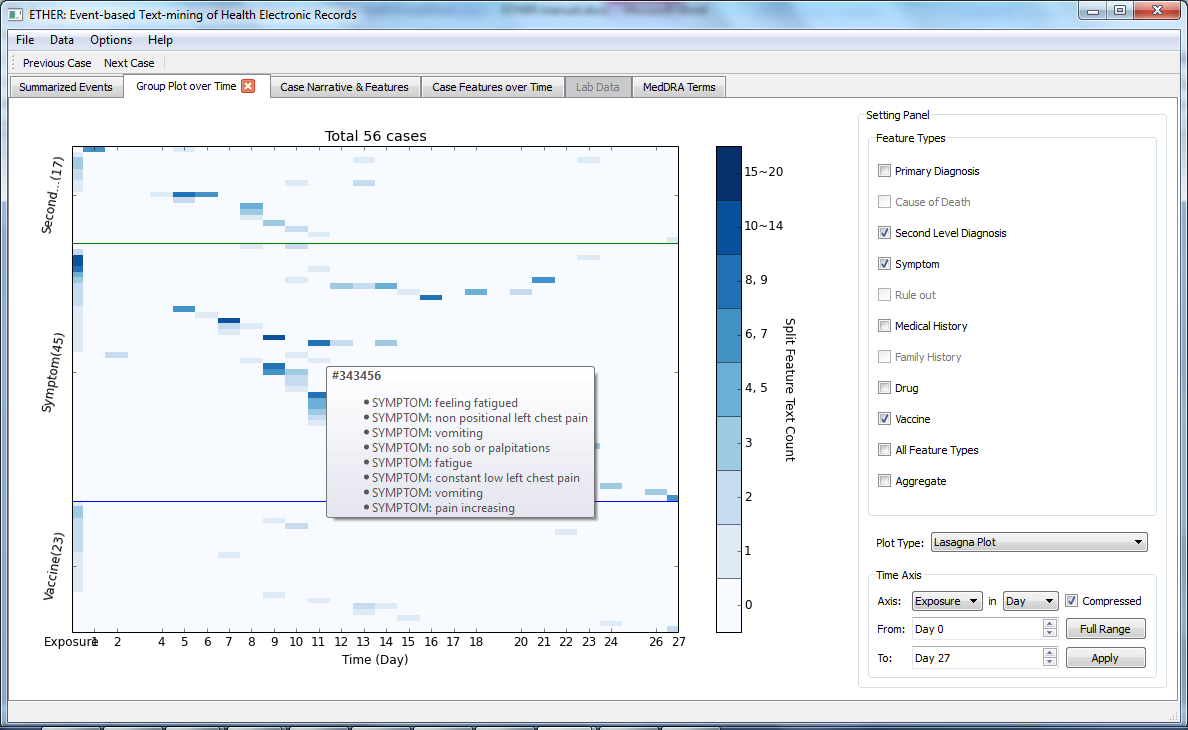


Figure 9: Lasagna plot

# Feature Count Plot

The Feature Count Plot shows the number of medical features over the current data set. When mouse is hovering over a colored block, a list of reports and their contributing is shown as a tooltip.

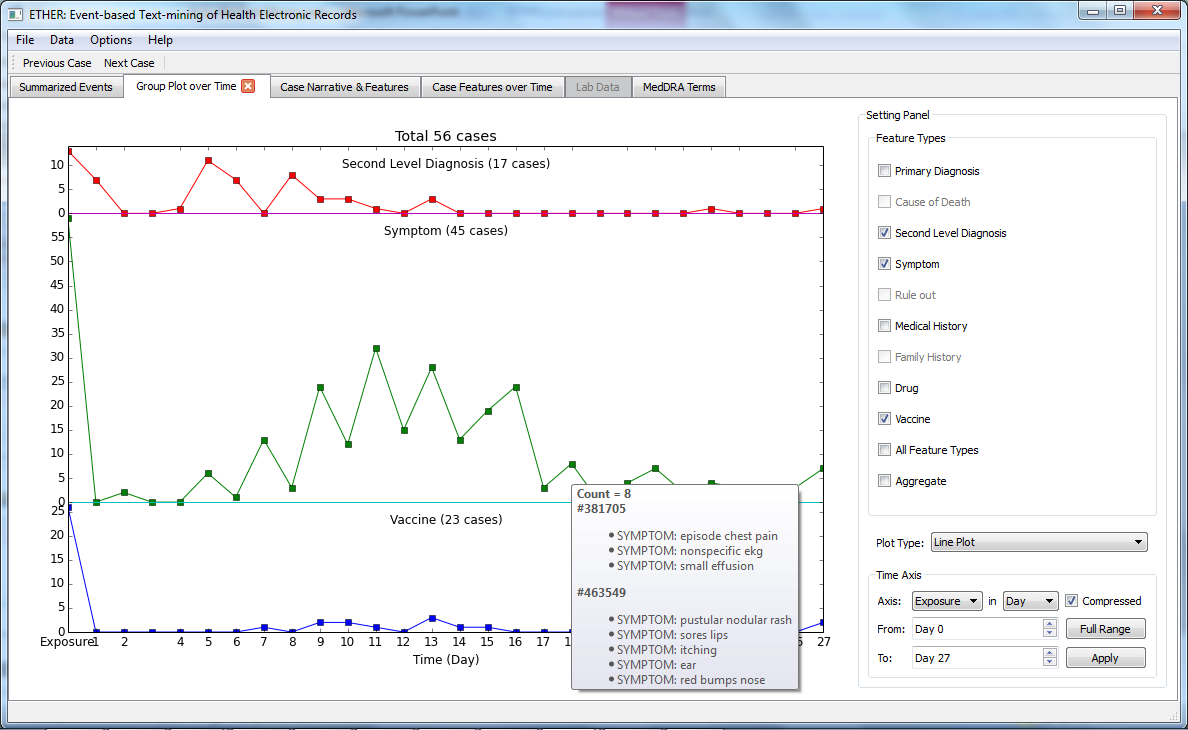


Figure 10: Feature Count Plot

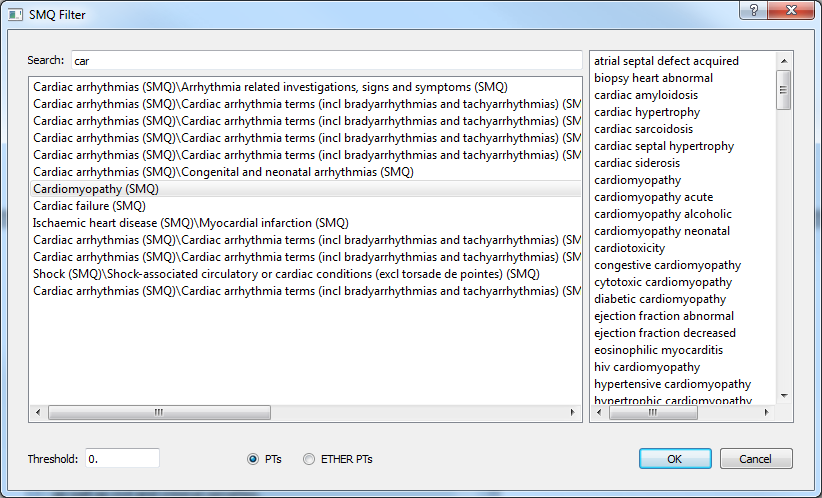
# Data Filtering

ETHER provides case filtering tools to select cases of interest based on extracted features and their time information.

# SMQ Tool

Choose *Data->Filter->SMQ* or use shortcut “Ctrl+S” to retrieve similar reports based on a Standardized MedDRA Query (SMQ), which is a list of Preferred Terms that are associated with a specific condition. The SMQ Case Builder window will appear, and a list of known SMQ’s should automatically be loaded. This list can be filtered by typing part of an SMQ name into the *Search* box. Click on an SMQ from the list to see all of the associated Preferred Terms to the right.

The *Threshold* text box is for the user to set a threshold for *Lin-similarity*, above which the reports with the reference SMQ are extracted. The user can choose to apply SMQ on either MedDRA PTs or PTs that are extracted and translated by ETHER with the radio buttons.

Figure 11: SMQ filter window

# Filter Builder

The filter builder can be opened through menu Data->Filter->Filter Builder. User can filter the current data set by feature type, text, and time. For advanced users, they can type the query directly in the text box on the top.

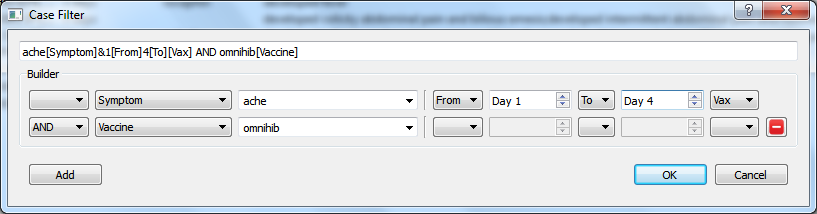


Figure 12: Filter builder

# System Preferences

ETHER provides a Preferences Dialog for users to set parameter for system appearance and behaviors. The preferences dialog can be opened through menu Options->Preferences.

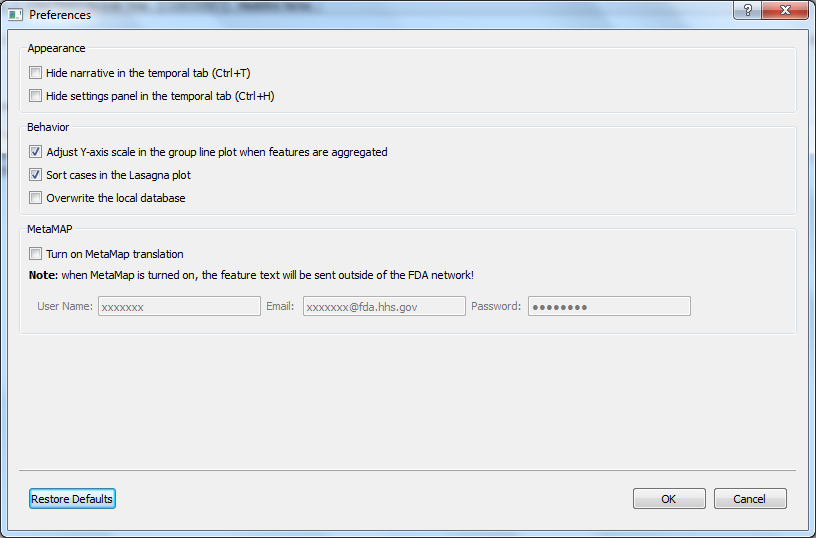


Figure 13: System Preferences Dialog