b. Convert the Horizontal Diagnosis Tree Dataset Into a Vertical Child-Parent[1] Structure

The second macro converts the diagnosis tree used by Tree Extraction in horizontal form, i.e. one record per diagnosis code and additional variables for each higher-level node of the tree, to a vertical form with one record per child-parent relationship which will be useful when identifying the next-higher node for any node which produces an alert. To illustrate, take a simple example from the current form of the tree based on the Multi-level Clinical Classifications Software (MLCCS), a product of the Agency for Healthcare Research and Quality's Healthcare Cost and Utilization Project (http://www.hcupus.ahrq.gov/toolssoftware/ccs/ccs.jsp). The ICD-9 diagnosis code for febrile seizures is 780.31, and the record in the diagnosis tree lookup dataset looks like this:

dx_codetype	Dx	mlccs5	mlccs4	mlccs3	mlccs2	mlccs1
09	780.31	780.31	06.04.02.00	06.04.02	06.04	06

When this horizontal record is converted into the vertical child-parent form, we have four records:

child	parent
780.31	06.04.02.00
06.04.02.00	06.04.02
06.04.02	06.04
06.04	06

Since future versions of the diagnosis tree could conceivably have a different number of levels than the current five, the conversion macro includes code to automatically determine the number of node variables in the horizontal tree and extract the nodes one by one into the vertical child-parent tree.

c. Convert the Child-Parent Tree Dataset Into dx-node Structure

In order to construct the alerts SAS lookup file to be used with the Data Freeze program, the third macro converts the child-parent tree into a different vertical form with one record per unique combination of diagnosis code and node. Taking the above example, we have the following five records:

dx	node
780.31	780.31
780.31	06.04.02.00
780.31	06.04.02
780.31	06.04
780.31	06

d. Automatically Identify Statistical Alerts Using Criteria Agreed Upon by the TreeScan Workgroup Prior to Analysis

The fourth macro selects the alerts for data freeze based on criteria that should be established by the FDA and TreeScan workgroup prior to analysis. First, primary alerts are selected from nodes meeting both of the following criteria: 1) p-value less than or equal to a maximum value; and 2) relative risk greater than or equal to a minimum value. These primary alerts are then compared to the child-parent

[1] The term "child-parent" here refers to tree structure terminology and not to a familial relationship between people.

but PBAs are time-consuming and resource-intensive, making them cost-prohibitive as a routine tool for investigation of statistical alerts arising from data mining.

The second aim of this activity fills the gap between broader investigation tools and detailed PBAs by creating re-usable programs to extract and retrieve patient-level case data for review by FDA and PRISM investigators. This type of review is not intended to validate the outcome or determine the validity of an alert but rather to determine whether further investigation is warranted. It is also not intended as an automatic first option but instead should be used only on a small subset of alerts, if at all, and then only after careful consideration of the circumstances by the surveillance team.

Interest in similar capabilities from outside the TreeScan workgroup led us to develop the Patient Episode Profile Retrieval (PEPR) as a self-contained macro requiring a single input dataset stored at the Data Partner site to identify patient episodes of interest. At minimum, the input dataset must contain a PatID variable and at least one date variable. Since the PEPR macro itself does not determine the patient episodes to be included, other methods must first be used to identify patient episodes of interest and complete any required sampling or sub-setting.

1. Security

The PEPR output datasets are based on the CDM tables with certain modifications made to protect patient privacy. In order to strike a balance between the need for robust patient-level case data and minimum-necessary data requirements that are fundamental to the distributed network model, we implemented both mandatory and optional security measures.

a. Mandatory

Pseudo-identifiers are automatically assigned to replace four identifiers – PatID, EncounterID, Provider, and Facility_code – in the output files, and crosswalks are saved at the local sites to allow translation to the original. Each pseudo-identifier is assigned using sequential numbering to assure uniqueness, but the original values are first randomly sorted to further mask the identifier. A random seed parameter is included to assure the randomization process can be reproduced, if necessary.
Additional care is taken in assignment of the PatID pseudo-identifier to account for situations in which a single patient may contribute more than one episode to an analysis. To distinguish between separate episodes for the same patient in the output datasets, particularly when using relative dates (described in the next section), the PatID pseudo-identifier contains two parts: the first identifies the patient, and the second identifies the episode. For example, the current Tree Extraction program defines incidence at the third level of the tree. If an alert occurred at a higher level, a patient could contribute more than one incident event to the alert, and if the events occurred on different days with the 1 – 56 days after exposure, we would not be able to distinguish between those events to assign appropriate relative dates and preserve data integrity unless we assign the pseudo-identifier to the patient-episode combination rather than to the patient alone.

Future TreeScan methods development will add multi-dose analysis, i.e. inclusion of doses beyond the first observed per patient. Once again, depending on the incidence definition and the spacing of doses, a patient could conceivably contribute the same or a similar incident event following each dose. Further, with an eye toward extending use of PEPR beyond TreeScan, we considered PBAs such as the PRISM evaluation of febrile seizures following influenza vaccination in young children.3 Incidence was defined as first observed in 42 days. Since small children are recommended to have two doses of the vaccine, a single patient could have been identified as a case once for each dose if a seizure followed each dose.

That evaluation also looked at PCV and DTaP-containing vaccines, and those additional exposures could have resulted in identification of separate adverse event episodes if the vaccines were administered on different days.

b. Optional

Date variables may be masked by calibrating values to a meaningful relative index variable specified by macro parameter. The relative index date value is subtracted from each CDM date variable so that the index date now has a value of 0 (SAS date value = Jan. 1, 1960), and then any other date now represents the number of days before or after that index date. For Tree Temporal Scan using a self-controlled analysis, the natural choice for relative index is the exposure date, which is then set to 0, and then all other dates represent how many days before or after exposure the encounter, drug dispensing, enrollment start, or enrollment end occurred. The format of the date variables is preserved but the identifiable information removed. Thus, the reviewer will not know the actual calendar date of an event, only the number of days before or after the relative index.

Exposure date	Original ADate value	New ADate value	Unformatted numeric value
10/01/2011	10/01/2011	01/01/1960	0
10/01/2011	10/10/2011	01/10/1960	10
04/15/2006	04/15/2006	01/01/1960	0
04/15/2006	03/15/2006	12/01/1959	-31

If this option is not selected, the original calendar date values are included in the output datasets. An activity involving chart review, for example, would need the original calendar dates in order to match chart data with electronic data.
The only date variable not included in this relative date option is the birth date, since application of the rule would result in representation of the patient's exact age in days on any given relative date. Instead, another option allows for the birth date value to be set to missing for all patients. This rule is optional since the birth date value is necessary for activities involving chart review.
As a compromise, an additional option allows the programmer to specify age groupings to more broadly categorize each patient's age at a selected index date variable.

The final option concerns two variables in the CDM that contain geographic information represented by the ZIP code: Zip in the Demographic table, representing the last known patient ZIP code, and Facility location in the Encounter table, valued with the first three digits of the facility ZIP. An optional macro parameter allows the values to be set to missing, converted to standard postal state abbreviation, or to retain their original value. If geographic clustering is suspected as a confounder, the postal state value may be used while still transmitting a lower level of specificity than the actual five- or three-digit ZIP values. The actual ZIP values would rarely – if ever – be required for anything less than full chart review and should only be used with extreme caution.

Any altered variables follow the same data type and format as the original CDM variables, but the variable name is changed by adding an underscore character "_" to the beginning of the original variable name. For example, PatID becomes _PatID, and ADate becomes _ADate. This convention serves as a reminder to anyone working with the datasets that those variables have been altered for security purposes. In order to run programs that were written to run on the CDM tables, the programmer only

CDM Table	Variable	Transformation
Enrollment	Enr_end	If prefixed with underscore, calculated as Enr_end – relative index date value, i.e. days +/- the relative index date. Otherwise, original value
Encounter, Diagnosis, Procedure	_EncounterID	Always masked; crosswalk retained in local folder
Encounter, Diagnosis, Procedure	ADate	If prefixed with underscore, calculated as ADate – relative index date value, i.e. days +/- the relative index date. Otherwise, original value
Encounter	DDate	If prefixed with underscore, calculated as DDate – relative index date value, i.e. days +/- the relative index date. Otherwise, original value
Encounter, Diagnosis, Procedure, State Vaccine	_Provider	Always masked; crosswalk retained in local folder
Encounter	Facility_location	If prefixed with underscore, set to missing or converted to postal state abbreviation, depending on macro parameter selection. Otherwise, original value
Encounter, Lab Results	_Facility_code	Always masked; crosswalk retained in local folder
Dispensing	RxDate	If prefixed with underscore, calculated as RxDate – relative index date value, i.e. days +/- the relative index date. Otherwise, original value
State Vaccine	SIIS	If prefixed with underscore, set to missing Otherwise, original value
State Vaccine	VaxDate	If prefixed with underscore, calculated as VaxDate – relative index date value, i.e. days +/- the relative index date. Otherwise, original value
Lab Results	Order_dt	If prefixed with underscore, calculated as Order_dt – relative index date value, i.e. days +/- the relative index date. Otherwise, original value
Lab Results	Lab_dt	If prefixed with underscore, calculated as Lab_dt – relative index date value, i.e. days +/- the relative index date. Otherwise, original value
Lab Results	Result_dt	If prefixed with underscore, calculated as Result_dt – relative index date value, i.e. days +/- the relative index date. Otherwise, original value
Vital Signs	Measure_date	If prefixed with underscore, calculated as Measure_date – relative index date value, i.e. days +/- the relative index date. Otherwise, original value
Death	DeathDate	If prefixed with underscore, calculated as DeathDate – relative index date value, i.e. days +/- the relative index date. Otherwise, original value

VII. ACKNOWLEDGEMENTS

We gratefully acknowledge the contributions of the following organizations and individuals:

Harvard Pilgrim Health Care Institute: Jeffrey Brown, Claudia Colonel-Moreno, Alison Kawai, Matt Lakoma, Jen Popovic, Lisa Ortendahl, Megan Reidy, Schyler Versteeg, Tiffany Woodworth, Diana Santiago, Robert Rosofsky

HealthCore: Bolarinwa Ekezue, Chunfu Liu, Fang Tian

Humana: Mano Selvan, Yunping Zhou

VIII. REFERENCES

1. Yih WK, Lieu TA, Kulldorff M, Martin D, McMahill-Walraven CN, Platt R, Selvam N, Selvan M, Lee GM, Nguyen M. Intussusception Risk after Rotavirus Vaccination in U.S. Infants. N Engl J Med, 2014, 370:503-12.
2. Yih WK, Zichittella LJ, Greene SK, et al. Evaluation of the Risk of Venous Thromboembolism After Gardisil Vaccination. Report to FDA; April 23, 2015. (Accessed at http://mini-sentinel.org/work_products/Assessments/Mini-Sentinel_PRISM_Gardasil-and-Venous-Thromboembolism-Report.pdf)
3. Kawai AT, Martin D, Kulldorff M, Li L, Cole DV, McMahill-Walraven CN, Selvam N, Selvan MS, Lee GM. Febrile Seizures After 2010–2011 Trivalent Inactivated Influenza Vaccine. Pediatrics, 2015-0635.
4. Kawai AT, Li L, Kulldorff M, et al. Mini-Sentinel CBER/PRISM Surveillance: Influenza Vaccines and Pregnancy Outcomes. Protocol for FDA; March 28, 2014. (Accessed at http://www.mini-sentinel.org/work_products/PRISM/Mini-Sentinel_PRISM_Influenza-Vaccines-and-Pregnancy-Outcomes-Protocol.pdf)
5. Kawai AT, Li L, Andrade SE, et al. Mini-Sentinel/CBER Protocol: Influenza Vaccines and Birth Outcomes. Protocol for FDA; December 5, 2014. (Accessed at http://www.mini-sentinel.org/work_products/PRISM/Mini-Sentinel_PRISM_Influenza-Vaccines-and-Birth-Outcomes-Protocol.pdf)
6. Yih WK, Zichittella LJ, Sandhu SK, et al. Accessing the Freshest Feasible Data for Conducting Active Influenza Vaccine Safety Surveillance. Report to FDA; April 8, 2015. (Accessed at http://www.mini-sentinel.org/work_products/PRISM/Mini-Sentinel_PRISM_Active-Influenza-Vaccine-Safety-Surveillance-Report.pdf)
7. TreeScan Extension (PRISM). Methods project for FDA; in progress. (Accessed at http://www.mini-sentinel.org/methods/methods_development/details.aspx?ID=1056)
8. TreeScan Power. Methods project for FDA; in progress. (Accessed at http://mini-sentinel.org/methods/methods_development/details.aspx?ID=1055)
9. Yih WK, Nguyen M, Maro JM, et al. Mini-Sentinel CBER/PRISM Surveillance: Pilot of Self-Controlled Tree-Temporal Scan Analysis for Gardasil Vaccine. Protocol for FDA; March 30, 2015. (Accessed at http://mini-sentinel.org/work_products/PRISM/Mini-Sentinel_PRISM_Pilot-Self-Controlled-Tree-Temporal-Scan-Analysis-Gardasil-Vaccine-Protocol.pdf)
10. McClure DL, Raebel MA, Yih WK, Shoaibi A, Mullersman JE, Anderson-Smits C, Glanz JM Mini-Sentinel methods: framework for assessment of positive results from signal refinement. Pharmacoepidemiology and Drug Safety. 2014;23:3-8.
11. Yih WK, Kulldorff M, Fireman BH, et al. Active surveillance for adverse events: the experience of the Vaccine Safety Datalink Project. Pediatrics. 2011;127:S54–64.
12. Kulldorff M. TreeScan User Guide, v1.1. November, 2014; 37-38. (Accessed at http://www.treescan.org/cgi-bin/treescan/register.pl/treescanv1.1-userguide.pdf?todo=process_userguide_download)
13. Klein NP, Fireman B, Yih WK, Lewis E, Kulldorff M, Ray P, Baxter R, Hambidge S, Nordin J, Naleway A, Belongia EA, Lieu T, Baggs J, Weintraub E, for the Vaccine Safety Datalink. Measles-mumps-rubella-varicella combination vaccine and the risk of febrile seizures. Pediatrics, 2010, 126, e1-8.