

How to use APT-DLD

(Automated Phenotyping Tool for identifying Developmental Language Disorder cases in health systems data)

Automated Phenotyping Tool for Identifying Developmental Language Disorder Cases in Health Systems Data (APT-DLD): a new research algorithm for deployment in large-scale electronic health record systems.

Journal of Speech, Language & Hearing Research, 2020

Courtney E. Walters Jr^{2, 4*}, Rachana Nitin^{1, 2*}, Katherine Margulis^{5, 7}, Olivia Boorom⁵, Daniel E. Gustavson^{2, 3}, Catherine T. Bush⁵, Lea K. Davis^{3, 6}, Jennifer E. Below^{3, 6}, Nancy J. Cox^{3, 6}, Stephen M. Camarata⁵, Reyna L. Gordon^{2, 3, 1}

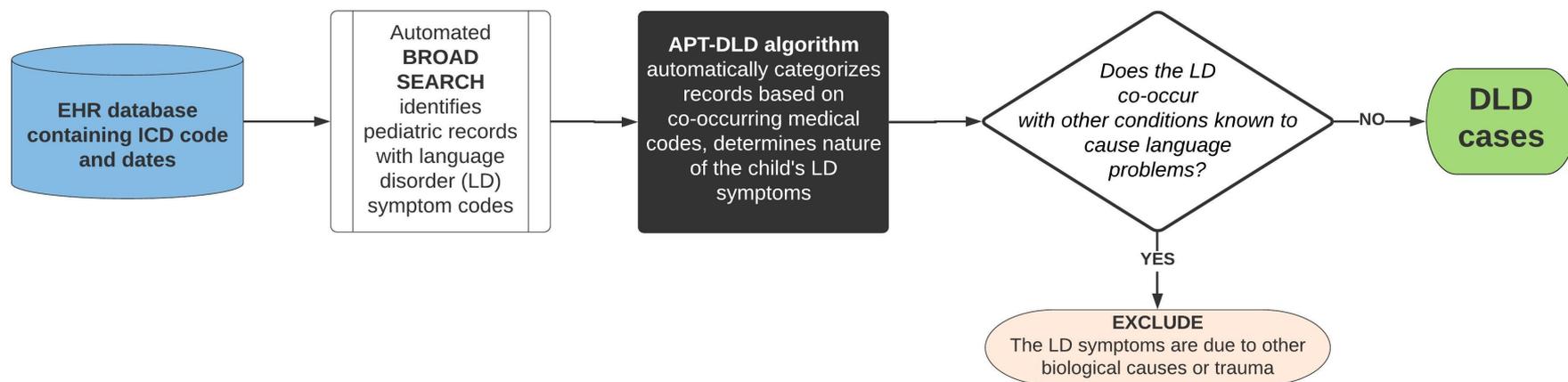
¹Vanderbilt Brain Institute, Vanderbilt University, ²Department of Otolaryngology, Vanderbilt University Medical Center, ³Vanderbilt Genetics Institute, ⁴Vanderbilt University Neuroscience Program, ⁵Department of Hearing and Speech Sciences, Vanderbilt University Medical Center, ⁶Department of Medicine, Vanderbilt University Medical Center, and ⁷Kennedy Krieger Institute.

*Authors CEW and RN have made equal contributions

Contact: reyna.gordon@alumni.usc.edu and rachana.nitin17@gmail.com

APT-DLD overview

Automated Phenotyping Tool for identifying DLD cases in health systems data (APT-DLD)



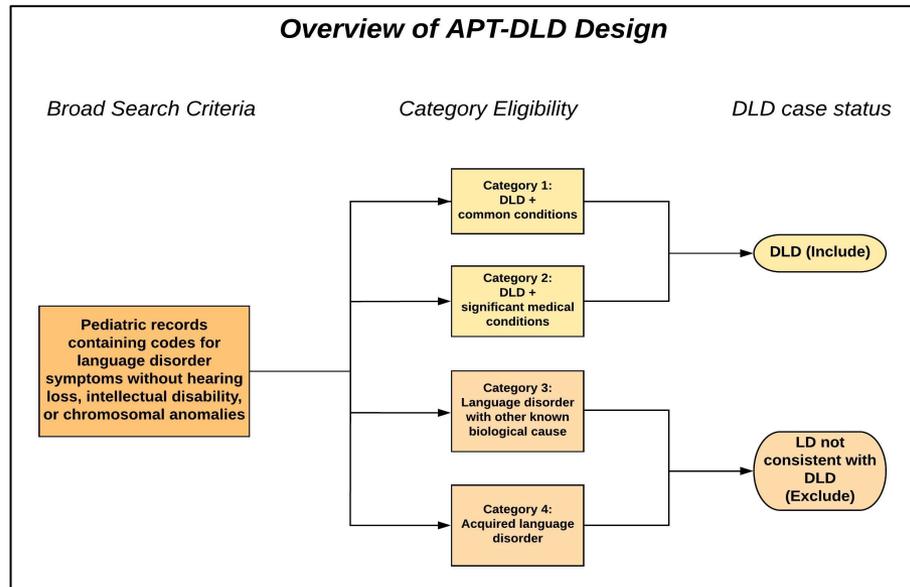
**APT-DLD is protected by a creative commons license
(CC BY-NC for Attribution-NonCommercial)**

Users must cite Walters, Nitin et al., 2020, JSLHR

July 28, 2020

APT-DLD

- Algorithm is fully automated and can be run in KNIME or R on either PC or Mac
- Uses filters to categorise electronic health records with language disorder into four categories
 - (see Developmental Language Disorder Manual Chart Review Rubric in Supplemental Information of Walters, Nitin, et al., 2020, *JSLHR* for details)



A. Formatting your data

Step 1a: Format your EHR data

- Your EHR data should be saved in a long format csv file.
- Columns in data file should appropriately be named the following:
 - GRID**
 - participant IDs
 - ICD_FLAG (optional)**
 - Indicates if the ICD_CODE is from the 9th or 10th edition
 - ICD_DATE**
 - MM/DD/YYYY format
 - ICD_CODE**
 - ICD_DESCRIPTION (optional)**
 - The description of the ICD_CODE

GRID	ICD_FLAG	ICD_DATE	ICD_CODE	ICD_DESCRIPTION
R876450495	10	9/12/2018	Q66.0	Congenital talipes equinovarus
R876450495	10	5/19/2018	R26.9	Unspecified abnormalities of gait and mobility
R876450495	9	3/18/2015	793.7	Nonspecific (abnormal) findings on radiological and other examination of musculoskeletal system
R876450495	10	5/11/2018	M21.162	Varus deformity; not elsewhere classified; left knee
R876450495	9	3/18/2015	879.8	Open wound(s) (multiple) of unspecified site(s); without mention of complication
R876450495	10	5/11/2018	Q66.0	Congenital talipes equinovarus
R876450495	10	6/23/2018	R29.91	Unspecified symptoms and signs involving the musculoskeletal system
R876450495	10	9/19/2018	M21.541	Acquired clubfoot; right foot
R876450495	10	5/11/2018	R62.50	Unspecified lack of expected normal physiological development in childhood
R876450495	10	7/13/2018	R29.91	Unspecified symptoms and signs involving the musculoskeletal system
R876450495	10	5/19/2018	Q66.0	Congenital talipes equinovarus
R876450495	10	6/23/2018	R62.50	Unspecified lack of expected normal physiological development in childhood
R876450495	10	9/4/2018	R26.9	Unspecified abnormalities of gait and mobility
R876450495	9	3/18/2015	E884.9	Other accidental fall from one level to another
R876450495	10	10/3/2018	M21.549	Acquired clubfoot; unspecified foot
R876450495	10	5/11/2018	R29.91	Unspecified symptoms and signs involving the musculoskeletal system
R876450495	10	9/4/2018	F80.2	Mixed receptive-expressive language disorder
R549274509	9	3/17/2015	959.9	Unspecified site injury
R549274509	10	9/4/2018	Q66.0	Congenital talipes equinovarus
R549274509	10	8/15/2018	R62.50	Unspecified lack of expected normal physiological development in childhood
R549274509	10	7/13/2018	Q66.0	Congenital talipes equinovarus
R549274509	9	3/18/2015	959.09	Injury of face and neck
R549274509	10	8/15/2018	Q66.0	Congenital talipes equinovarus
R549274509	10	9/4/2018	R46.89	Other symptoms and signs involving appearance and behavior
R549274509	10	7/13/2018	R62.50	Unspecified lack of expected normal physiological development in childhood
R549274509	10	9/19/2018	M21.542	Acquired clubfoot; left foot
R549274509	10	5/19/2018	R29.91	Unspecified symptoms and signs involving the musculoskeletal system
R549274509	10	5/11/2018	M21.161	Varus deformity; not elsewhere classified; right knee
R549274509	10	5/19/2018	R62.50	Unspecified lack of expected normal physiological development in childhood
R549274509	10	9/4/2018	R62.50	Unspecified lack of expected normal physiological development in childhood
R549274509	9	3/19/2011	750	Tongue tie
R549274509	9	1/3/2016	734	Flat foot
R549274509	9	10/18/2011	752.49	Other anomalies of cervix; vagina; and external female genitalia
R549274509	9	10/15/2013	788.41	Urinary frequency
R549274509	9	3/27/2011	V67.00	Follow-up examination; following surgery; unspecified
R549274509	9	2/25/2011	750	Tongue tie
R549274509	9	2/25/2011	787.21	Dysphagia; oral phase

Step 1b: Format your Demographic data

GRID	GENDER	RACE_ASIAN	RACE_AFRICAN_AMERICAN	RACE_HISPANIC	RACE_NATIVE_AMERICAN	RACE_CAUCASIAN	RACE_OTHER	RACE_UNKNOWN	ETHNICITY	DOB	DECEASED
C542347800	M	0	0	1	0	0	0	0	0 HL	11/24/2012	null
C367452080	F	0	0	0	0	0	1	0	0 NH	4/7/2002	null
C336450582	M	0	1	0	0	0	0	0	0 NH	3/29/2002	null
C745784780	M	0	0	0	0	0	1	0	0 NH	4/26/2011	null
C745345769	M	0	0	1	0	0	0	0	0 HL	12/6/2010	null
C123430869	F	0	0	0	0	0	1	0	0 NH	7/1/2002	null
C123432356	F	0	0	0	0	0	1	0	0 NH	5/24/2002	null
C167872356	F	0	0	0	0	0	1	0	0 NH	3/12/2012	null
C576478800	M	0	0	0	0	0	1	0	0 NH	8/25/1997	null
C569812309	F	0	0	0	0	0	1	0	0 NH	1/16/2003	null

- Your demographic data should be in wide format csv or excel file
- The most important columns in the data are
 - GRID
 - Date of Birth (will be used to eliminate non-pediatric patients with language codes)
 - All other information is optional

B. APT-DLD Using R

You will need

- R and R studio (any version would do, latest versions are advised)
R: <https://www.r-project.org/>
R studio: <https://rstudio.com/>
- EHR data file (.csv format; shown in Part A, step 1)
- Two .zip folders from PheKB
 - Broad Search_prepareforAPT-DLD.zip
 - APT-DLD_implementationR.zip

Step 1: Download PheKB files

- Access APT-DLD at <https://phekb.org/phenotype/developmental-language-disorder>
- Or:
 - Open up the PheKB website at: <https://phekb.org>
 - Search for *Developmental Language Disorder* in the search bar

The screenshot shows the PheKB website homepage. The header includes the PheKB logo and tagline 'a knowledgebase for discovering phenotypes from electronic medical records'. Navigation links for 'Home', 'Phenotypes', 'Resources', and 'Contact Us' are present. A search bar is located in the top right corner, highlighted with a red box. A red arrow points from the text 'Search for Developmental Language Disorder in the search bar' to this search bar. The main content area features a central diagram titled 'What is the Phenotype KnowledgeBase?' with four interconnected nodes: 'Access Validated Phenotype Algorithms', 'Validate existing phenotype algorithms on your EMR', 'Collaborate on Phenotype Algorithms', and 'Share Validated Phenotype Algorithms'. Text on the right explains the importance of health data and the collaborative nature of PheKB. A sidebar on the right lists 'Most Recent Phenotypes' including 'Digital Rectal Exam', 'bone scan utilization', 'PHEMA BPH (Benign Prostatic Hyperplasia) cases', 'Systemic lupus erythematosus (SLE)', and 'Multimodal Analgesia'. The footer contains information about using the website and cross-site collaboration.

Step 1: Download PheKB files

- Access the phenotype *Developmental Language Disorder* in PheKB to download:
 - BroadSearch_prepareforAPT-DLD.zip
 - APT-DLD_implementationR.zip

The screenshot displays the PheKB interface for phenotype ID 797. At the top, there are three tabs: 'Phenotype' (selected), 'Edit', and 'Data Dictionaries'. Below the tabs, the following information is presented:

- Phenotype ID: 797**
- Status:** Final
Do Not List on the Collaboration Phenotypes List
- Type of Phenotype:** Disease or Syndrome, Other Trait
- Data Modalities and Methods Used:** ICD 10 Codes, ICD 9 Codes
- Contact Author:** Reyna Gordon
- Files:** A list of four files is shown, each with a document icon:
 - BroadSearch_prepareforAPT-DLD.zip
 - APT-DLD_implementationR.zip
 - APT-DLD_implementationKNIME.zip
 - APT-DLD_filters.zip

The 'Files' section is highlighted with a red border in the original image.

Step 2: Broad Search

- The “Broad Search” identifies an initial pool of pediatric patient records with language codes, upon which APT-DLD can be run in Step 3
 - We have provided R code for the Broad Search, but some researchers can use their own database mining methods
 - The list of contingencies can be found in the “Lang codes.csv” and “Broad search exclusion codes.csv” in the BroadSearch_preparforAPT-DLD.zip file
- Set up R and R studio for appropriate computing device
- Open BroadSearch_prepareforAPT-DLD.zip and load Broad Search_Public.R into R studio
- Run the script as directed.
- The output of the Broad Search serves as the input for Step 3

Step 3: Run APT-DLD in R

- Load the APT-DLD_Algorithm_public.R script
- Run the script as directed
- The output will be 2 files (both files are lists of GRIDs/pt IDs that fall into one of the 4 categories)
 - Inclusion cases (Cat 1 + Cat 2 cases)
 - Exclusion cases (Cat 3 + Cat 4 cases)

C. APT-DLD Using KNIME

You will need

- R and R studio (any version would do, latest versions are advised, for the broad search)
R: <https://www.r-project.org/>
R studio: <https://rstudio.com/>
- KNIME <https://www.knime.com/downloads/download-knime>
- EHR data file (.csv format; shown in Part A, step 1)
- Three .zip folders from PheKB
 - BroadSearch_prepareforAPT-DLD.zip
 - APT-DLD_implementationKNIME.zip
 - APT-DLD_filters.zip

Step 1: Download KNIME

- Download KNIME software on appropriate computing device from:
<https://www.knime.com/downloads/download-knime>
- Download the KNIME Analytics Platform for Windows (installer package)
- The algorithm was built using the KNIME version 3.7.2, however we are confident that it will function on newer KNIME builds as well

Download KNIME Analytics Platform

1 Register for Help & Updates 2 **Download KNIME** 3 Get Started

Download the latest KNIME Analytics Platform for Windows, Linux, and Mac OS X.

KNIME 3.7.2

Find out **What's New in the new release** [here](#).

The KNIME Analytics Platform version is intended for end users and provides everything needed to immediately begin using KNIME as well as extend KNIME with extension packages developed by others.

Windows		
KNIME Analytics Platform for Windows (installer) <i>The installer adds an icon to the desktop and suggests suitable memory settings</i>	64 Bit (441.03 MB) 32 Bit (437.42 MB)	
KNIME Analytics Platform for Windows (self-extracting archive) <i>The self-extracting archive only creates a folder holding the KNIME installation</i>	64 Bit (444.58 MB) 32 Bit (441.15 MB)	
KNIME Analytics Platform for Windows (zip archive)	64 Bit (529.54 MB) 32 Bit (525.59 MB)	

Linux		
KNIME Analytics Platform for Linux	64 Bit (554.2 MB)	

Mac		
KNIME Analytics Platform for Mac OSX (10.11 and above)	64 Bit (522.98 MB)	

Step 2: Download APT-DLD materials from PheKB

- Access APT-DLD at <https://phekb.org/phenotype/developmental-language-disorder>
- Or:
 - Go to <https://phekb.org>
 - Search for *Developmental Language Disorder* in the search bar

The screenshot shows the PheKB website homepage. The header includes the PheKB logo and tagline 'a knowledgebase for discovering phenotypes from electronic medical records'. A search bar is located in the top right corner, highlighted with a red box. A red arrow points from the text 'Search for Developmental Language Disorder in the search bar' to this search bar. The main content area features a central diagram titled 'What is the Phenotype KnowledgeBase?' with four interconnected nodes: 'Access Validated Phenotype Algorithms', 'Share Validated Phenotype Algorithms', 'Collaborate on Phenotype Algorithms', and 'Validate existing phenotype algorithms on your EMR'. Text on the right explains that health data is becoming an important source for clinical and genomic research, and that PheKB is a collaborative environment for building and validating electronic algorithms. A 'Most Recent Phenotypes' sidebar lists items like 'Digital Rectal Exam' and 'bone scan utilization'. The footer contains the text: 'On PheKB you can: View existing algorithms; Enter or create new algorithms; Collaborate with others to create or review algorithms; View implementation details for existing algorithms'.

Step 2: Download PheKB files

- Access the phenotype *Developmental Language Disorder* in PheKB to download:

- BroadSearch_prepareforAPT-DLD.zip
- APT-DLD_filters.zip
- APT-DLD_implementationKNIME.zip

The screenshot shows the PheKB interface for phenotype ID 797. At the top, there are three tabs: "Phenotype" (selected), "Edit", and "Data Dictionaries". Below the tabs, the following information is displayed:

- Phenotype ID: 797**
- Status:** Final
Do Not List on the Collaboration Phenotypes List
- Type of Phenotype:** Disease or Syndrome
Other Trait
- Data Modalities and Methods Used:** ICD 10 Codes
ICD 9 Codes
- Contact Author:** Reyna Gordon
- Files:** (highlighted with a red box)
 - BroadSearch_prepareforAPT-DLD.zip
 - APT-DLD_implementationR.zip
 - APT-DLD_implementationKNIME.zip
 - APT-DLD_filters.zip

Step 3: Perform the Broad Search- this requires R

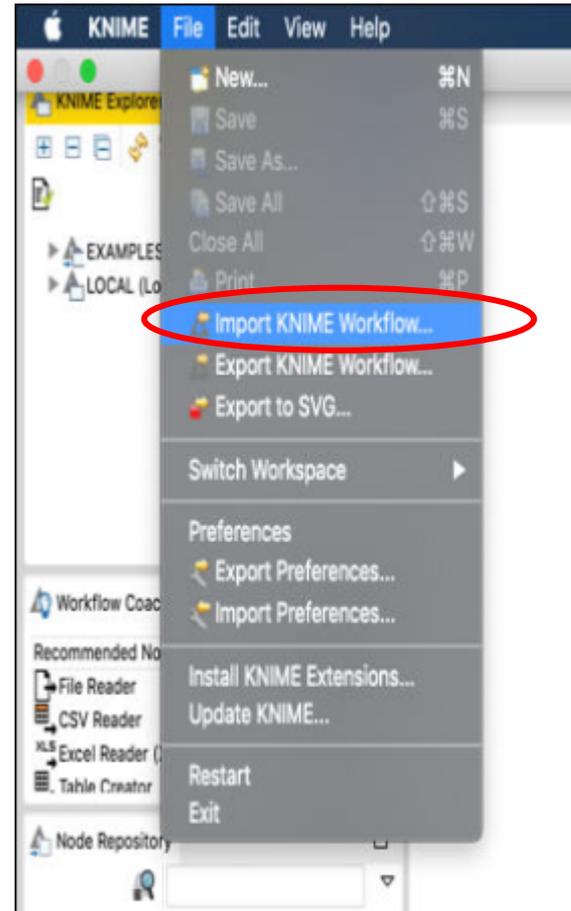
- The broad search consists of primary criteria to select those EHRs with valid language disorder codes
- Unzip BroadSearch_prepareforAPT-DLD.zip
 - Broad Search_Public.R
 - Lang Codes
 - Broad Search Exclusion Codes
- Run the Broad Search.R file in R studio
- The final file created will be the input for APT-DLD algorithm

The screenshot shows a web interface with three tabs: "Phenotype", "Edit", and "Data Dictionaries". The "Phenotype" tab is active, displaying the following information:

- Phenotype ID: 797**
- Status:** Final
Do Not List on the Collaboration Phenotypes List
- Type of Phenotype:** Disease or Syndrome, Other Trait
- Data Modalities and Methods Used:** ICD 10 Codes, ICD 9 Codes
- Contact Author:** Reyna Gordon
- Files:**
 - BroadSearch_prepareforAPT-DLD.zip (highlighted with a red box)
 - APT-DLD_implementationR.zip
 - APT-DLD_implementationKNIME.zip
 - APT-DLD_filters.zip

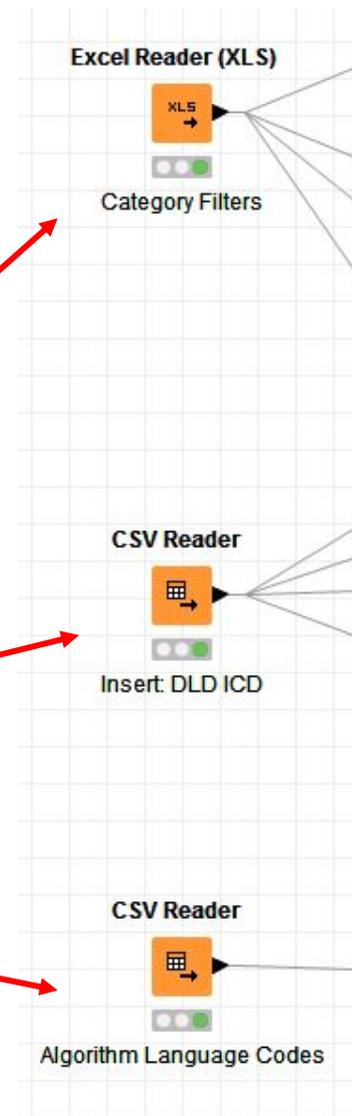
Step 4: Set up KNIME

- Unzip
APT-DLD_implementationKNIME.zip
- Open KNIME software
and import APT-DLD.knwf
from the unzipped folder



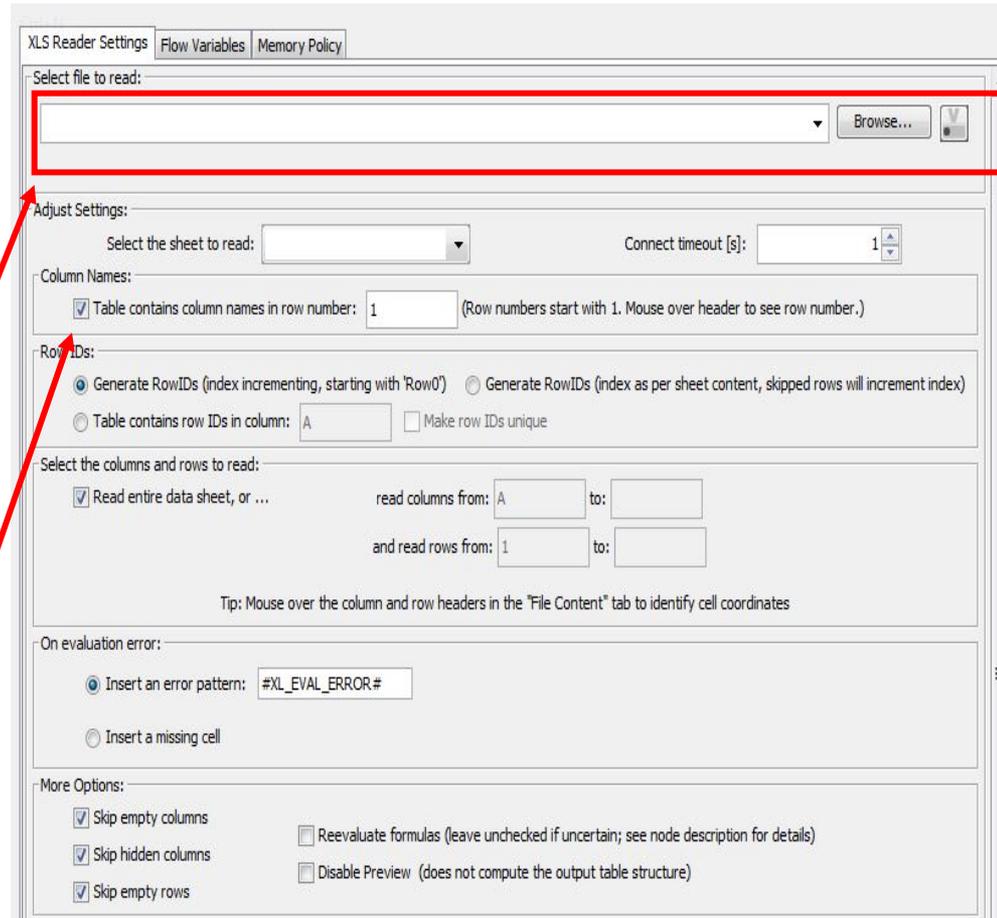
Step 5: Load required files into APT-DLD algorithm in KNIME

- Open APT-DLD in KNIME
- Unzip APT-DLD_filters.zip
- There are three files to load into APT-DLD
 - *Category_filters.xlsx*
 - *Your DLD ICD data csv file generated from the broad search*
 - *Algorithm Language Codes.csv*



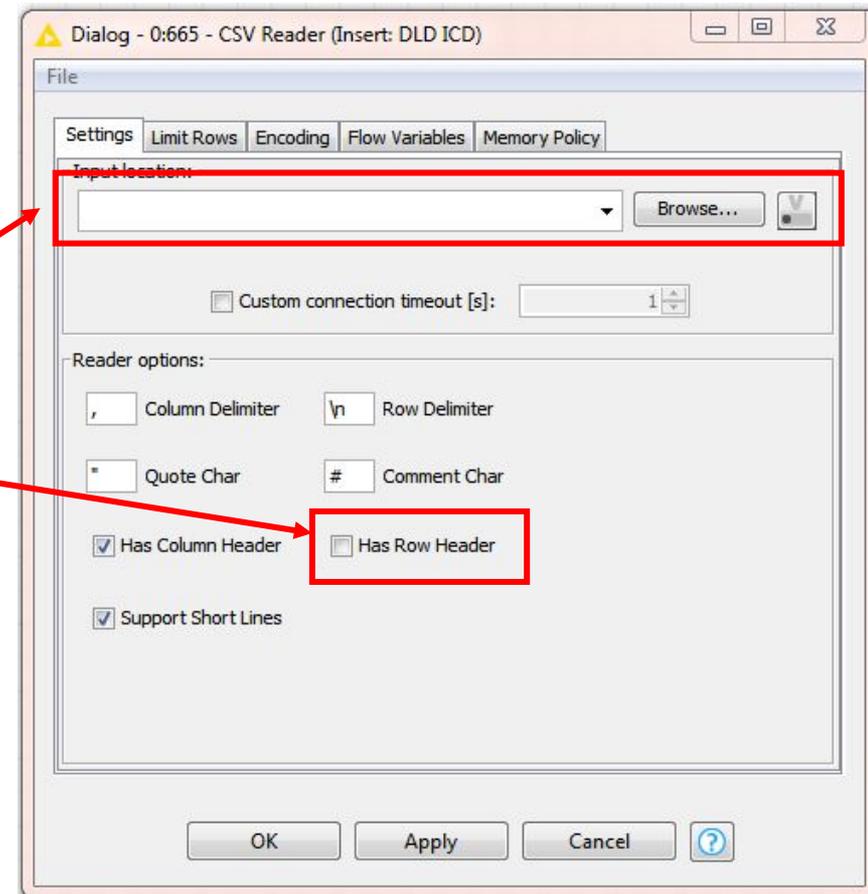
Step 6a: Loading xlsx file into APT-DLD

- Double click on the previously highlighted “Category Filters” XLS reader node. This allows you to configure the data
- Browse and enter the *Category_filters.xlsx* file into the node
- Check the box for “Table contains column name in row number 1”
- Click OK to load the file



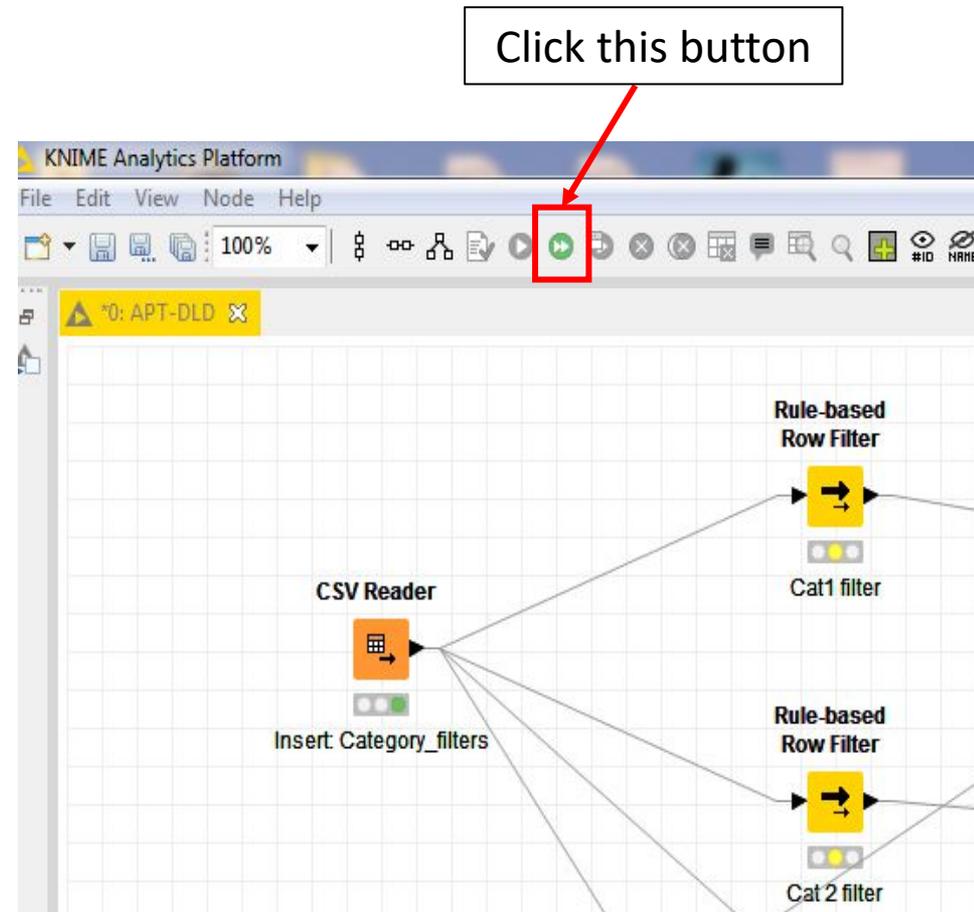
Step 6b: Loading csv files into APT-DLD

- Double click on each of the previously highlighted csv reader nodes. This allows you to configure the data
- Browse and enter in appropriate csv files to read for each node
(Algorithm Language Codes.csv into the Algorithm Language Codes csv reader, and the csv output from the Broad search into the DLD ICD node)
- Uncheck the box for “Has Row Header”
- Check the box for “Has Column Header”
- Check box for “Support Short Lines”
- Click OK to load the file



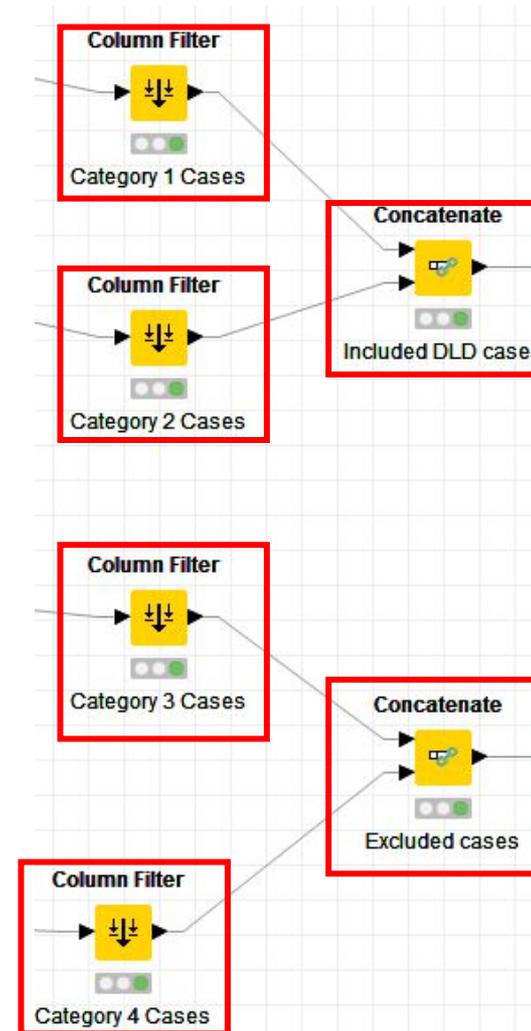
Step 7: Deploy APT-DLD

- Once all inputs are entered, each executable node will be ready to run as indicated by the **yellow** dots.
 - A red node indicates the node has not been configured
- Click “Execute all executable nodes” button to run the algorithm.
- When finished, all nodes will turn **green**.



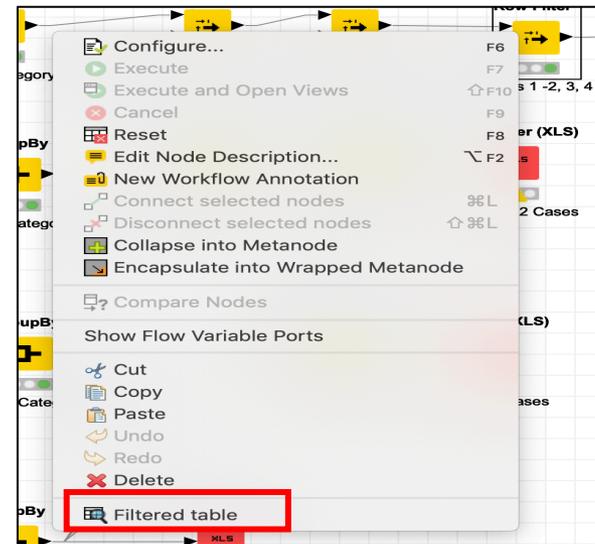
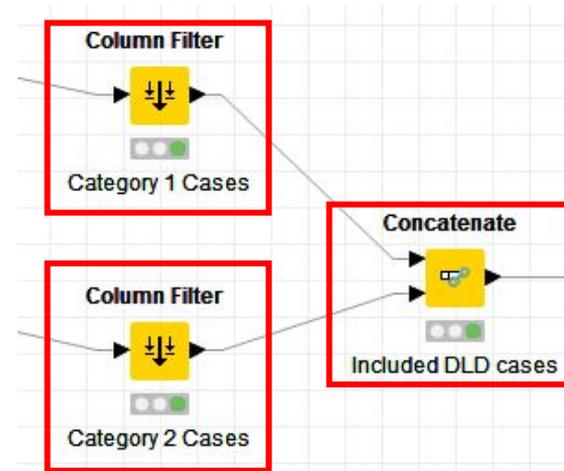
Step 8: Check results of APT-DLD

- The algorithm will use the category filters and classify your EHR into 4 categories (Category 1-4)
- Categories 1+2 are concatenated into the “Included DLD cases” node
- Categories 3+4 are concatenated into the “Excluded cases” node
- The output files will be the “included” and “excluded” cases.



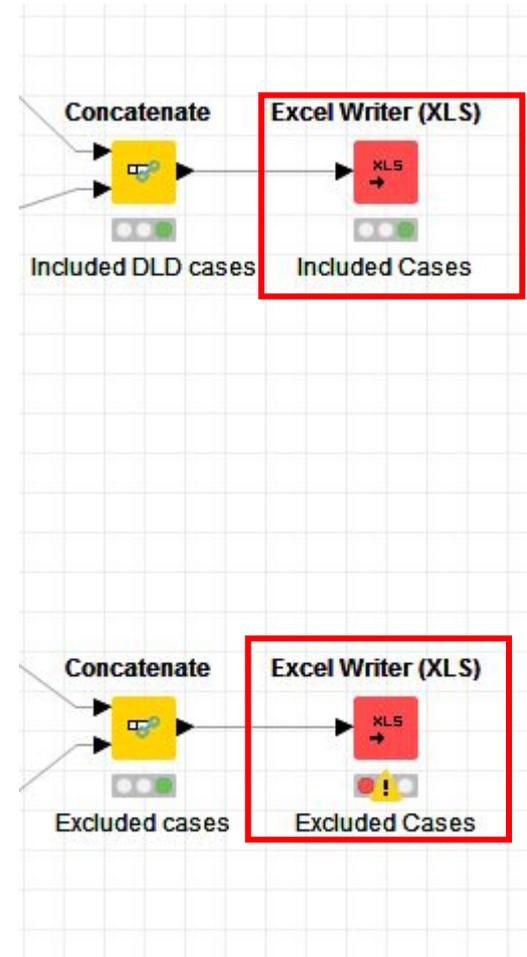
Step 8: Check results of APT-DLD

- Right click on any of the 4 Category Cases (highlighted) or on the included or excluded cases nodes and click on “Filtered table” to view results.
- The results will be displayed as a pop-up KNIME window with the table of GRIDS in the category selected.



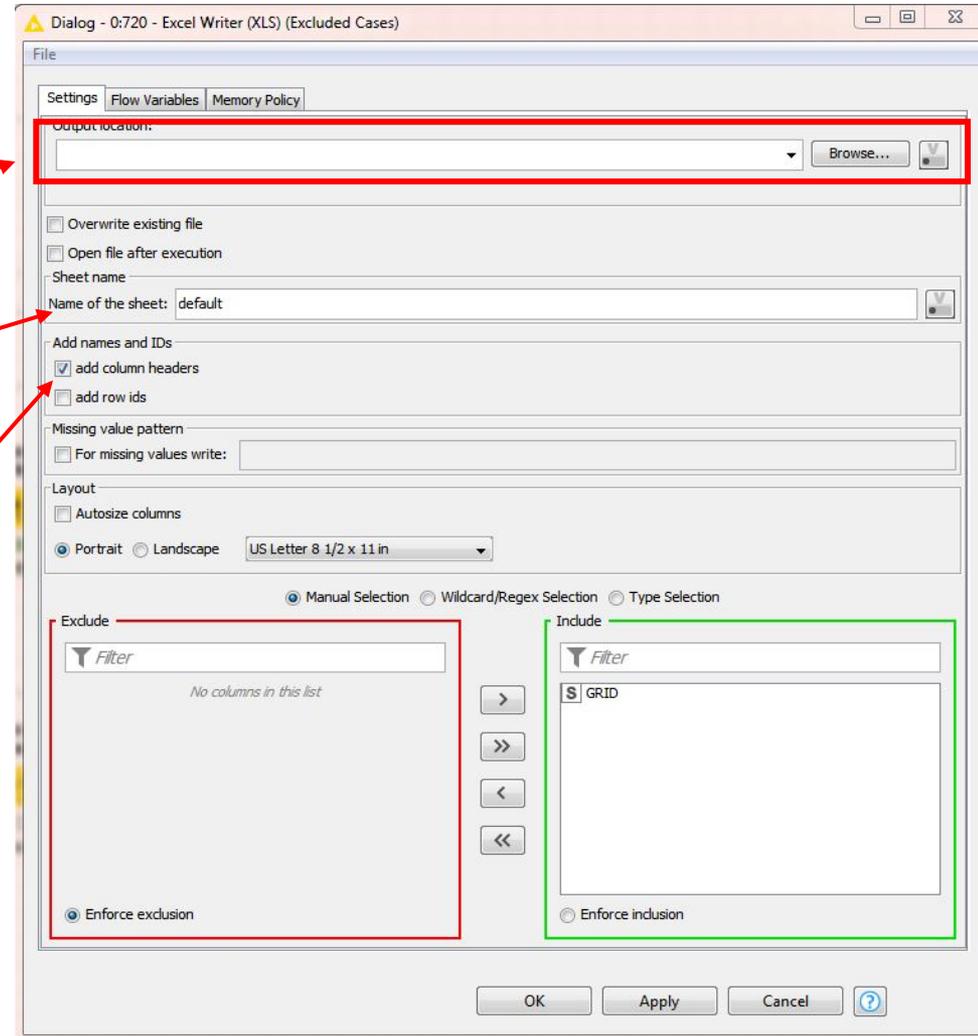
Step 9: Exporting results as Excel files

- The included and excluded cases can be exported to excel files.
- To export results, double click on the “Excel Writer (XLS) node” to configure them



Step 9: Exporting results as Excel files

- Browse and enter in appropriate location to export excel files to.
- Name the sheet in excel file (optional).
- Check the box for “add column headers”.
- Click OK.



APT-DLD users must agree to cite:

Walters, Nitin, Margulis, Boorom, Gustavson, Bush, Davis, Below, Cox, Camarata, & Gordon (2020). *Automated Phenotyping Tool for Identifying Developmental Language Disorder Cases in Health Systems Data (APT-DLD): a new research algorithm for deployment in large-scale electronic health record systems*. *Journal of Speech, Language & Hearing Research*.

Need help or want to discuss further?

Contact: reyna.gordon@alumni.usc.edu and
rachana.nitin17@gmail.com

FIN