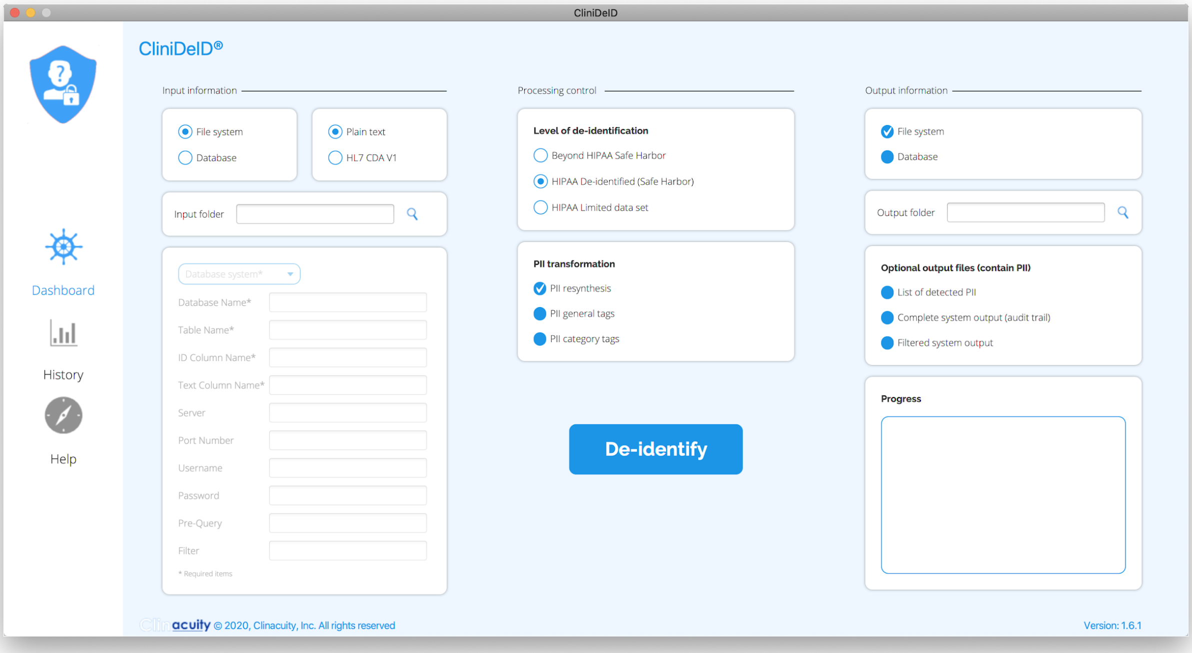
CliniDeID

Using the Application

Summary

The adoption of Electronic Health Record (EHR) systems is growing at a fast pace in the U.S., and this growth results in very large quantities of patient clinical data becoming available in electronic format, with tremendous potential, but also equally growing concern for patient confidentiality breaches. Secondary use of clinical data is essential to fulfill the potentials for high quality healthcare, improved healthcare management, and effective clinical research. Clinacuity, Inc. proposes a new system to automatically de-identify clinical notes found in the EHR, to then improve the availability of clinical text for secondary uses, as well as ameliorate the protection of patient data confidentiality: CliniDeID®.



Input information

◉ File system
○ Database

◉ Plain text
○ HL7 CDA V1

Input folder

Database system*

Database Name*
Table Name*
ID Column Name*
Text Column Name*
Server
Port Number
Username
Password
Pre-Query
Filter

* Required items

Processing control

Level of de-identification

○ Beyond HIPAA Safe Harbor
◉ HIPAA De-identified (Safe Harbor)
○ HIPAA Limited data set

PII transformation

☑ PII resynthesis
● PII general tags
● PII category tags

Output information

☑ File system
● Database

Output folder

Optional output files (contain PII)

● List of detected PII
● Complete system output (audit trail)
● Filtered system output

Progress

acuity © 2020, Clinacuity, Inc. All rights reserved

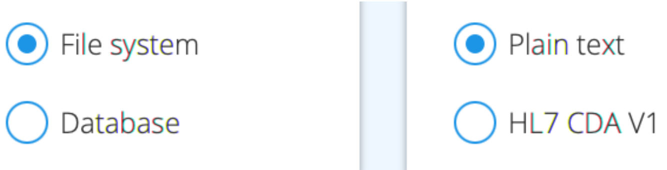
Version: 1.6.1

Version 1.9.0, Apr 29, 2023

How to Use

Choosing Inputs

CliniDeID can obtain the data to process either from a directory of files or from a database query. If the 'File System' option is chosen (the default) then click the 🔍 button next to the 'Input folder' box to navigate to your directory and click "Open" in the file navigator. If 'Plain Text' is selected then only .txt files will be processed while only .xml files are processed with the 'HL7 CDA V1' option. You may also enter the directory paths directly in the text fields on the application.

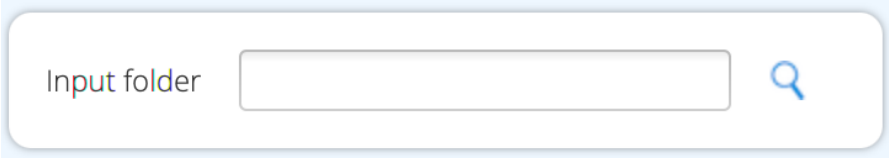


● File system

○ Database

● Plain text

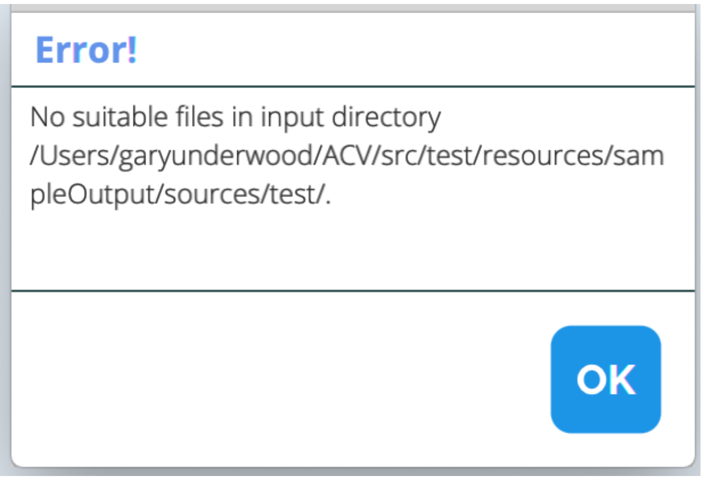
○ HL7 CDA V1



Input folder

In the data folder is a directory named *sampleTextFromI2B2* with text files from the I2B2 2014 De-identification track that can be used as sample data.

Note: The system only processes files with the proper extension (.txt for plain text and .xml for HL7 CDA). Other files will be ignored, with a warning displayed. No sub-directories are allowed in the input directory. The input directory requires read access. The application will display an error, similar to the one on the right, if any of these criteria are not met.

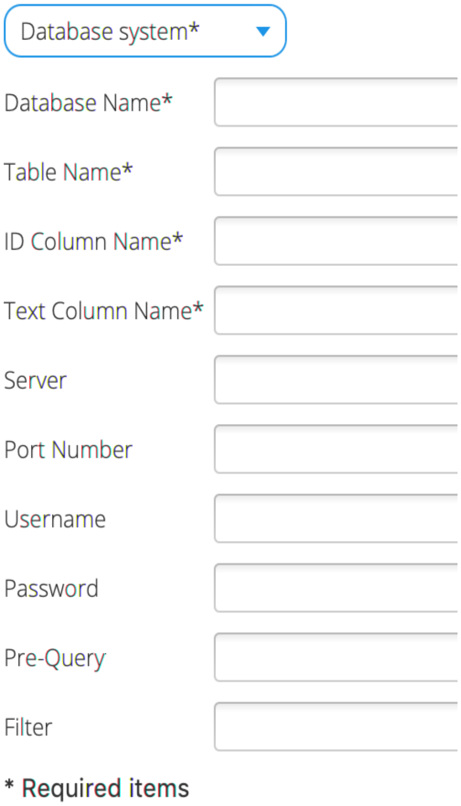


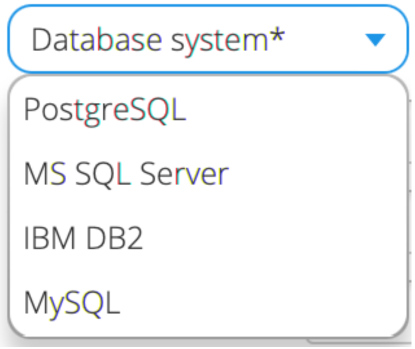
Error!

No suitable files in input directory /Users/garyunderwood/ACV/src/test/resources/sampleOutput/sources/test/.

Instead of files, a database query can also be used to provide data for CliniDeID. Select 'Database' and additional fields are required.

First, select your database type from the dropdown list and fill in the other fields appropriately. Contact your database administrator for these values. The system will perform a Select query on the database of the form 'SELECT ID_Column_Name, Text_Column_Name FROM Table_Name'. The 'Pre-Query' field is used if there is a need to run a SQL statement before the SELECT statement, for example to change schemas or views. The 'Filter' field is used to restrict which rows from the table will be used, for example 'WHERE ID_Column_Name = 12345'. The 'Filter' field's value is added to the end of the 'SELECT' statement.





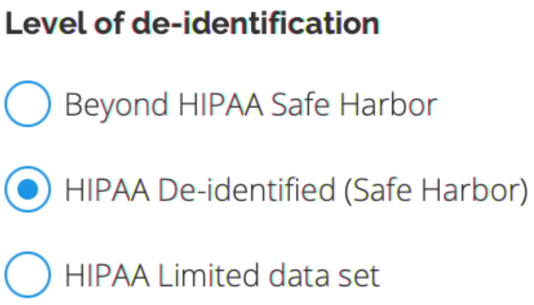
Choosing Outputs

CliniDeID can output to files, to a PostgreSQL database or to both. Even if only database outputs are chosen, an output folder must still be selected for a ProcessedDocumentsList.txt file that contains a list of the documents and when they were processed. If database outputs are chosen then the PostgreSQL database must already be created and its server running. See the README file for instructions on the one time installation and setup of the PostgreSQL database as well as information about how the database is organized. The names of output files will be the same as the input files (with different extensions) or the value of ID_Column_Name depending on input source. The output directory requires write access or an error message will be displayed.

Options

Level of De-identification

CliniDeID supports 3 levels of de-identification in the Options menu.



Level of de-identification

○ Beyond HIPAA Safe Harbor

● HIPAA De-identified (Safe Harbor)

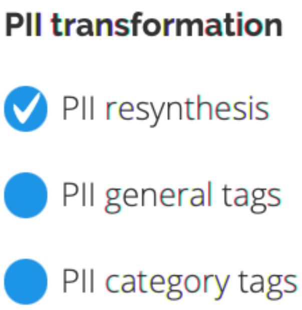
○ HIPAA Limited data set

- Beyond HIPAA Safe Harbor - in addition to Safe Harbor categories, it identifies all ages, health care providers, states, countries, professions, years, and full zip codes.
- HIPAA Safe Harbor - all HIPAA Safe Harbor PII (Personally Identifiable Information) categories. This is the default.
- Limited Data Set - Safe Harbor but no ages, no dates, and no zip codes.

Custom - The custom option allows for choosing exactly which PII subtypes for CliniDeID to process. Custom settings can be saved (they are plain xml files) and shared with other users for consistent use.

Choosing PII Transformations

There are three transformations the application provides which controls what CliniDeID will replace PII with. More than one may be selected. The application will not run unless at least one transformation or optional output (described below) is selected.



PII transformation

☑ PII resynthesis

● PII general tags

● PII category tags

1. PII resynthesis - The original document de-identified and resynthesized with realistic PII surrogates.

2. PII general tags - The original document de-identified with PII replaced by generic [***PII***] tags.

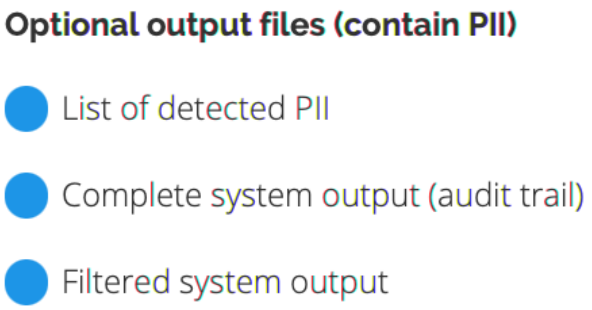
3. PII category tags - The original document de-identified with PII replaced by generic [***PII Category***] tags (Category includes name, location, etc.).

If HL7 CDA is selected for input, then PII transformation outputs will be too.

Choosing Optional Output Files

1. List of detected PII* - A list of the PII detected with their category and subcategory (patient name, date, ...)

2. Complete system output (audit trail) * - XMI format file with all annotations found by the system (including pre-processing and PII) for auditing purposes (file output only)



Optional output files (contain PII)

● List of detected PII

● Complete system output (audit trail)

● Filtered system output

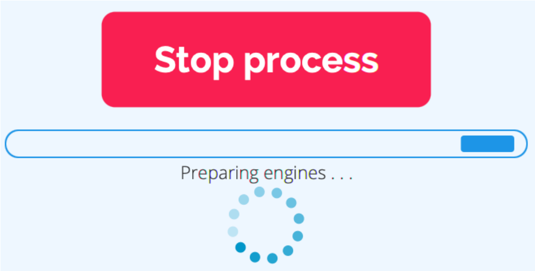
3. Filtered system output * - XMI format file with only the final PII Annotations (file output only)

* These output formats may contain PII.

Running CliniDeID

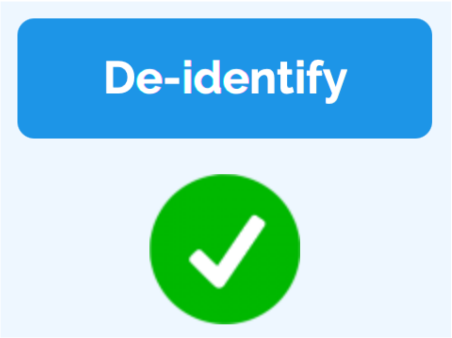
Once output types and level of de-identification have been selected (or if you want to return all outputs by default) click the “Run Deid” button on the bottom right of the application. If there was an error caused by user input the application will display it immediately; otherwise the progress bar will show a loading animation. It may take several minutes to prepare the engines.

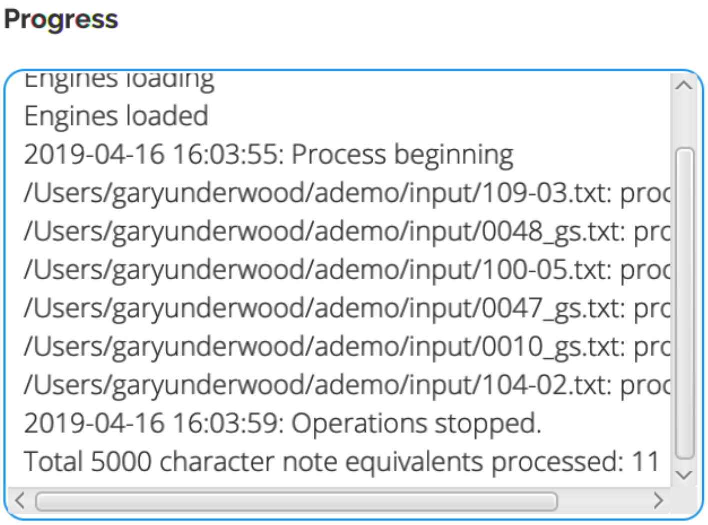
To cancel processing click Stop process.. There will be a delay while the system cleans up resources.



Preparing engines . . .

While the application is running the Progress box will show updates. As each file is finished it will be listed. When the application is finished running success will be indicated and the progress box will show the total of 5000 character note equivalents which may be used for billing. If the application displayed an error message see the "Troubleshooting Errors" section.



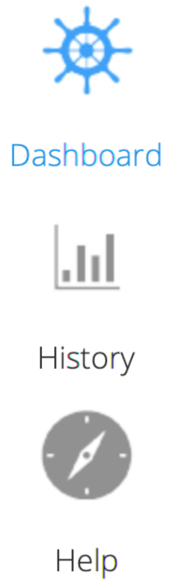


```
Engines loading
Engines loaded
2019-04-16 16:03:55: Process beginning
/Users/garyunderwood/ademo/input/109-03.txt: proc
/Users/garyunderwood/ademo/input/0048_gs.txt: prc
/Users/garyunderwood/ademo/input/100-05.txt: proc
/Users/garyunderwood/ademo/input/0047_gs.txt: prc
/Users/garyunderwood/ademo/input/0010_gs.txt: prc
/Users/garyunderwood/ademo/input/104-02.txt: proc
2019-04-16 16:03:59: Operations stopped.
Total 5000 character note equivalents processed: 11
```

Viewing Output

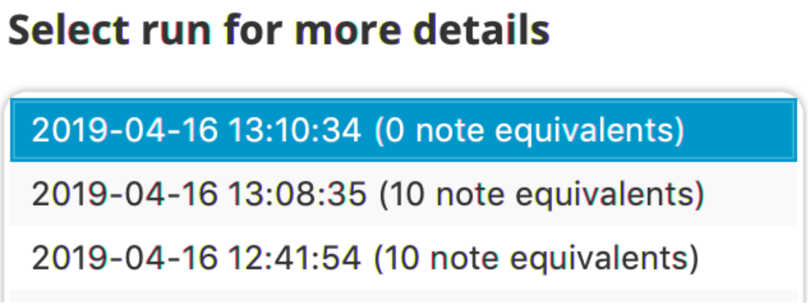
To view output files, navigate to the output directory and open the files with your preferred text, xml, or xmi editor. If output is sent to the PostgreSQL database then view the output by querying the appropriate tables.

History Information



The History information is available by selecting it from the sidebar on the left.

A list of the runs will be displayed, with the most recent at the top. Each entry is the time of the run in year-month-day hour:minute:second format together with how many note equivalents were completed in that run. This does not include processing done from other folders on this machine or on other machines. The complete count of note equivalents processed is available at https://deid.clinacuity.com. Selecting a run will show the details of that run in the box on the right. The details shown are those displayed in the Progress box during execution.



Select run for more details

2019-04-16 13:10:34 (0 note equivalents)
2019-04-16 13:08:35 (10 note equivalents)
2019-04-16 12:41:54 (10 note equivalents)

Command Line Operation

CliniDeID can also be run from the command line with the Windows .bat file or Mac/Unix sh script file “runCliniDeIDcommandLine”. See the Readme file in the CliniDeID folder for details.

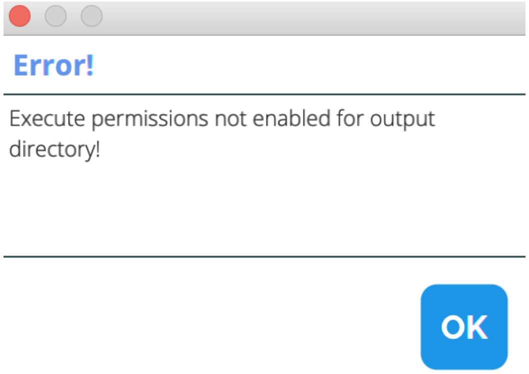
Troubleshooting Errors

Note that during execution CliniDeID connects to the license server through port 443 (standard for https).

Error Types

The system currently recognizes two error types: Errors caused by faulty user input, and errors caused by a system malfunction.

1. User Input Errors - If the application displays an error related to the input/output directories or input files you must fix the issue to run CliniDeID. You must ensure that the input directory has read access, only files with the .txt extension, and no sub-directories. The output directory must allow write access. Once the errors have been corrected try running the application again.
2. System Errors - If the application displays the generic error message shown to the right there was an issue with CliniDeID itself. To view the file click the “Log” button at the top of the application. You may also locate the file within the CliniDeID directory in a subfolder named log. The log file contains information that helps developers find and fix issues with the application. Log files do contain processed filenames so if your filenames contain PII (e.g. patient identifier) then the log will contain PII. Otherwise, there is no document data in the log that could be PII.



Error!

Execute permissions not enabled for output directory!

License and Copyright Information

For more information visit the Clinacuity website at https://.clinacuity.com/ In the folder *data/license* are subfolders for each of the libraries used by CliniDeID and their license and/or copyright information. Here is a short list of libraries and software used directly by CliniDeID:

Liblinear: R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A Library for Large Linear Classification, Journal of Machine Learning Research 9(2008), 1871-1874. Software available at http://www.csie.ntu.edu.tw/~cjlin/liblinear

Mira, jaxb-api, xmlbeans, xom, jython-standalone, OpenJDK OpenJFX, mallet, spring, aws-java-sdk-ec2, snakeyaml, python, keras, tensorflow, rnn (BiLSTM-CNN-CRF)

And from APACHE:

Maven, UIMA, uimaj-core, UimaFit, ctakes-type-system, ctakes-utils, jdom2, commons-cli,opennlp-tools, regex annotator, log4j, log4j-core, log4j-api, log4j-slf4j-impl, log4j-1.2-api, log4j-jcl, clear-TK, lucene, postgreSQL, Open Sans font, Roboto font.

Professions list: Government of Western Australia, Dept of Training and Workforce Development, https://www2.jobsandskills.wa.gov.au/career-exploration/Occupations/Pages/OccupationsA-Z.asp

DB2 jdbc drivers from IBM.

SqlServer jdbc driver from Microsoft.

MySql and Oracle jdbc drivers from Oracle.

Font: Raleway from Rodrigo Fuenzalida, sourcecodepro from Adobe, fira from Mozilla.