# Open Tools for Data De-identification Handout

## Definitions from NCDS Data Glossary

[**Artificial Intelligence (AI):**](https://www.nnlm.gov/guides/data-glossary/artificial-intelligence)AI consists of computer programs that are typically built to adaptively update and enhance their own performance over time. They are used to process, analyze, and recognize patterns in large datasets, and they use those patterns to get better at completing tasks or solving problems.

[**Machine Learning (ML):**](https://www.nnlm.gov/guides/data-glossary/machine-learning) Machine Learning is a type of Artificial Intelligence. Machine Learning involves sophisticated algorithms which can be trained to sort information, identify patterns, and make predictions within large sets of data.

[**Natural Language Processing (NLP):**](https://www.nnlm.gov/guides/data-glossary/natural-language-processing) deals with how computers understand, process, and manipulate human languages. It can involve things like interpreting the semantic meaning of language, translating between human languages, or recognizing patterns in human languages.

## Your Notes

## NLM Scrubber – Main Talking Points​

* NLM Scrubber is an openly available tool that can deidentify clinical text data and make it HIPPA compliant. ​
* Administration doesn’t have to buy or pay a subscription to use this tool.  ​

### **Argument for Need ​**

* NIH DMSP strongly encourages researchers to share their data, NLM Scrubber can help researchers be more compliant and competitive on grants. ​
* Engaging in good data practices will increase the impact of their research​.
* Deidentifying clinical text data makes it easier to share data and further protects patient privacy.
* Sharing data is good for science! It aids in reproducibility, transparency and helps researchers comply with funder mandates.

### **Possible Questions​**

* How much does it cost?
  + It’s freely available online. Anyone can download and use it.
* Is it secure/safe?
  + Yes! NLM Scrubber is run locally on your computer. It’s not connected to the cloud.
* How does it work?
  + Developed by the National Library of Medicine, it uses AI and Natural Language Processing to find personally identifying information and replace it with generic terms.
* Does it really work?
  + Yes, but a human oversight is needed. ​

## Resources to Get Started

[NLM Scrubber Website](https://lhncbc.nlm.nih.gov/scrubber/index.html)

[NLM Scrubber User Manual](https://data.lhncbc.nlm.nih.gov/public/scrubber/files/user_manual/windows/user_manual_v.19.0411W.pdf)

[NLM Scrubber Product Guide](https://www.nnlm.gov/BBK86) (downloads automatically)

[From the CDC: What is Personally Identifying Information?](https://www.cdc.gov/nchs/training/confidentiality/training/page581.html)

[Data Curation Primers](https://datacurationnetwork.org/outputs/data-curation-primers/)

[Human Participant Data Primer](https://github.com/DataCurationNetwork/data-primers/blob/main/Human%20Participants%20Data%20Essentials%20Data%20Curation%20Primer/human-participants-data-essentials-data-curation-primer.md)

## LibGuide Examples

[University of North Texas Health Science Center: Research Data Management Resources & Free Courses](https://libguides.unthsc.edu/research-data-management/resources)

[John Hopkins Sheridan Libraries: Protecting Human Subject Identifiers](https://guides.library.jhu.edu/protecting_identifiers/software)

## Scholarly Articles on NLM Scrubber and De-identification

[A Comparative Analysis of Speed and Accuracy for Three Off-the-Shelf De-Identification Tools. AMIA Joint Summits on Translational Science proceedings](https://pubmed.ncbi.nlm.nih.gov/32477643/) – from AMIA Joint Summits on Translational Science

[[The pattern of name tokens in narrative clinical text and a comparison of five systems for redacting them](https://doi.org/10.1136/amiajnl-2013-001689)](https://doi.org/10.1136/amiajnl-2013-001689) – *Journal of the American Medical Informatics Association (JAMIA)*

[De-identification of Address, Date, and Alphanumeric Identifiers in Narrative Clinical Reports](https://pubmed.ncbi.nlm.nih.gov/25954383) – from AMIA Symposium, 2014

[Exploring Freely Available Data Tools to Support Open Data and Open Science](https://pubmed.ncbi.nlm.nih.gov/38883700/) – *Journal of Hospital Librarianship*

[[Flexible data anonymization using ARX—Current status and challenges ahead](https://doi.org/10.1002/spe.2812)](https://doi.org/10.1002/spe.2812) – *Journal of Software: Practice and Experience*

[Protected Health Information filter (Philter): accurately and securely de-identifying free-text clinical notes](https://pubmed.ncbi.nlm.nih.gov/32337372/) – npj | Digital Medicine

[Evaluation of Automated Public De-Identification Tools on a Corpus of Radiology Reports](https://pubmed.ncbi.nlm.nih.gov/33937843/) – *Radiology: Artificial intelligence*

[An atomic approach to the design and implementation of a research data warehouse](https://pubmed.ncbi.nlm.nih.gov/34613409/) – *Journal of the American Medical Informatics Association (JAMIA)*

## Other De-identification Tools

* [CliniDeID](https://github.com/Clinacuity/CliniDeID)
* [ARX Data Anonymization](https://arx.deidentifier.org/)
* [OpenNLP](https://opennlp.apache.org/)
* [DICOM Library](https://www.dicomlibrary.com/) – for images
* [Philter](https://github.com/BCHSI/philter-ucsf)
* [MITRE Identity Scrubber Toolkit (MIST)](https://mist-deid.sourceforge.net)
* [DataShield](https://www.datashield.org/)
* [NeuroNER](http://neuroner.com/)