Use of Machine-Learning Software to Categorize Oncotype DX Data

By

Keven Gomez

Honors Thesis
UNC Eshelman School of Pharmacy
University of North Carolina at Chapel Hill

March 6, 2022

Approved:

Megan Roberts, Ph.D.

Abstract

Introduction:

Precision medicine provides a method of individualizing and optimizing therapy with the end goal of increasing efficacy, benefits, and success of therapeutic treatments for each patient. Because implementation of precision medicine requires the interpretation of several patient-characteristics, compiling all this information into one database that can be utilized efficiently is a challenge especially when data is found in unstructured data fields. This study looked into the utilization of machine-learning software to abstract and structure Oncotype DX testing results, a genomic test result that is often reported in text-heavy unstructured clinic notes, into a format that can be merged with larger structured databases.

Methods:

This study created a cohort of breast cancer patients within the UNC Health Care System who were eligible for Oncotype DX testing from 2015-2016. Oncotype DX testing results were manually abstracted for all the patients and each were labeled as being low, intermediate, high or no score based on their risk level. A subset of patients were used to create an algorithm for CLARK, the machine learning software, which was tasked with labeling these same patients on its own. The algorithm was run on another subset of patients, known as the evaluation corpus, and the labeling accuracy was compared to the manual abstraction.

Results:

1190 patients were utilized to test CLARK. 297 patients were used to create the algorithm and 893 patients were separated into the evaluation corpus. After running the gold standard algorithm on the evaluation corpus, it was determined that CLARK was 83.3% accurate in its labeling of Oncotype DX eligible patients compared to manual abstraction.

Conclusions:

CLARK provides a useful, accurate and efficient means of classifying large volumes of patients into structured data sets. While not 100% accurate, this study does highlight the efficiency it provides and the utility of improving machine-learning methods for future use.

1. Introduction:

Precision medicine is defined by the National Institutes of Health as “an emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle for each person.”[1] It does away with the one-size-fits-all approach and utilizes patient-specific characteristics, to individualize and optimize therapy with the end goal of increasing efficacy, benefits and success of therapeutic treatment for each patient.[2,3]. Improvements in the collection of precision medicine data has led to data sets of increasing volume; Big data are defined as “large data sets that may be analyzed computationally to reveal patterns, trends, and associations, especially relating to human behavior and interactions.”[3] The implementation of big data analytics has been trailing behind due to missing key data elements that would allow for researchers to better measure the impact of precision medicine on the quality of care. The full benefits of precision medicine can only be understood if there are methods to efficiently access, organize and analyze key parameters.

This is a multicomponent project, focused on compiling, and organizing genomic test results, specifically the Oncotype DX (ODX) test results, to then link it with other structured clinical and claims data. This manuscript will primarily focus on the compilation and organization of ODX test results. Created in 2004, the ODX test examines gene expression of twenty-one tumor-related genes in estrogen-receptor positive, HER-2 negative breast cancer. The test reports a 10-year risk of recurrence score, ranging from 0-100, where people can be arranged into low, intermediate, and high risk of recurrence categories based on their score.[5,6] The risk score aids in predicting the likelihood of 10-year recurrence and aids in predicting whether the patient would benefit from chemotherapy. Individuals with a low risk score are recommended to forgo chemotherapy, whereas individuals with a high risk score are recommended to receive chemotherapy. Currently these data are considered unstructured data, meaning ODX test results are often found in text-heavy fields (e.g., clinic notes) that cannot be easily organized. Normally, to gather ODX results requires one to manually search individual patient profiles by their medical record number (MRN) in the Electronic Health Record (known as EPIC in the UNC Health Care System). This becomes time-consuming, the more patients involved. This project addressed this issue and aimed to identify the most efficient way to abstract and organize this information into a database using a machine-learning classification software known as the Clinical Annotation Research Kit or CLARK.[7,8] CLARK is a user-friendly interface that combines algorithms known as computable phenotypes with natural language processing to utilize key words or phrases not available in structured data in identify and define the patient cohorts.

These patient cohorts were defined by recurrence score ranges: low, intermediate, high or no score if the patient did not receive an ODX test.

The Cancer Information and Population Health Resource (CIPHR) is a cancer registry linked to insurance claims associated with cancer care for both public (Medicare and Medicaid) and private payers (state BCBS) in North Carolina.[9] There also exists the Clinical Data Warehouse for Health (CDW-H), which stores structured electronic health record data, including clinical and billing data, from hospitals at UNC Health.[10] The second part of this project will involve linking data abstracted and organized by CLARK to these two databases to then assesses the completeness and agreement of data across these different sources. The end goal is to create a complete, structured dataset that can be used to evaluate the costs, access, and quality of breast cancer care delivery at the population level.

Reiterating that precision medicine requires account genomic data, patient and environmental factors to best be used for cancer care, the formatting and structuring of ODX results is important. A structured data set containing ODX results will allow for results to be combined with these other data sets easily in order to study its impact on cancer health outcomes. Further, this study may provide a process by which other unstructured precision medicine data can be abstracted from the electronic health record. This manuscript details the portion of this project testing new methods of data extraction that can be used instead of time-consuming manual health record abstraction. It is hypothesized that the use of machine-learning classification software such as CLARK will provide a much more efficient means of organizing unstructured data such as ODX testing results with acceptable accuracy.

2. Methods

2.1: Population definitions:

The first step of this project required the enumeration of our breast cancer cohort. Within UNC Health, the name of the database where clinical data is stored is known as the Carolina Clinical Data Warehouse for Health (CDW-H). Additionally, there is the Cancer Information and Population Health Resource (CIPHR) that stores data on healthcare claims used in cancer treatment or office visits linked with the state cancer registry.[9,10] When linked, these two databases can identify breast cancer patients within UNC Health who have Medicare, Medicaid or Blue Cross Blue Shield health insurance. CIPHR and CDW-H databases were initially linked to identify 5,593 patients with breast cancer who were diagnosed with breast cancer from 2015-

2016. The inclusion criteria for this study were imposed on the CIPHR data and included confirmed diagnosis of stage I and II, ER-positive, HER-2-negative breast cancer and eligibility for ODX testing. Additionally, within CDWH, patients had to have had at least 1 visit at any facility within UNC Health, two weeks prior to date of diagnosis and at least 1 visit up to 12 months after. This study was IRB approved (IRB-19-2261).

2.2: Identifying Oncotype DX-eligible patients:

The initial cohort of breast cancer patients of 5,593 patients needed to be refined to only contain breast cancer patients eligible for ODX testing. Because receipt of Oncotype DX testing is often an unstructured data field in the electronic health record, a program known as the Electronic Medical Record Search Engine (EMERSE) was utilized to refine this cohort to only patients who possibly qualified for Oncotype DX testing. EMERSE is a general-purpose term-searching engine, created in the University of Michigan in 2005, that allows users to search free text in clinic notes like a google search.[11] It differs from CLARK in that it lacks the natural language processing to improve its algorithm on its own and relies solely on the user; EMERSE can only report what MRNs are associated with your keywords. For the portion of this project, EMERSE's purpose was to omit MRNs that did not have any mention of Oncotype DX testing as a way to identify individuals who may have had Oncotype DX testing. Testing of several keywords such as: oncotype, ODX, oncotype dx, Er(+), Her-2 (-), genetic testing, was required to identify the keyword that best reported potential Oncotype DX testing. For a test of accuracy, a sample size of MRN's were taken for each keyword utilized to determine which keywords accurately caught the most individuals that had ODX testing discussed in their notes in a manual search in EPIC. Based on the sample sizes taken, it was found that the single keyword 'oncotype' was the most accurate. The key term 'Oncotype' was used to perform an EMERSE search refining our study cohort 1198 candidates that had ODX testing discussed with them and was received or opted out of testing.

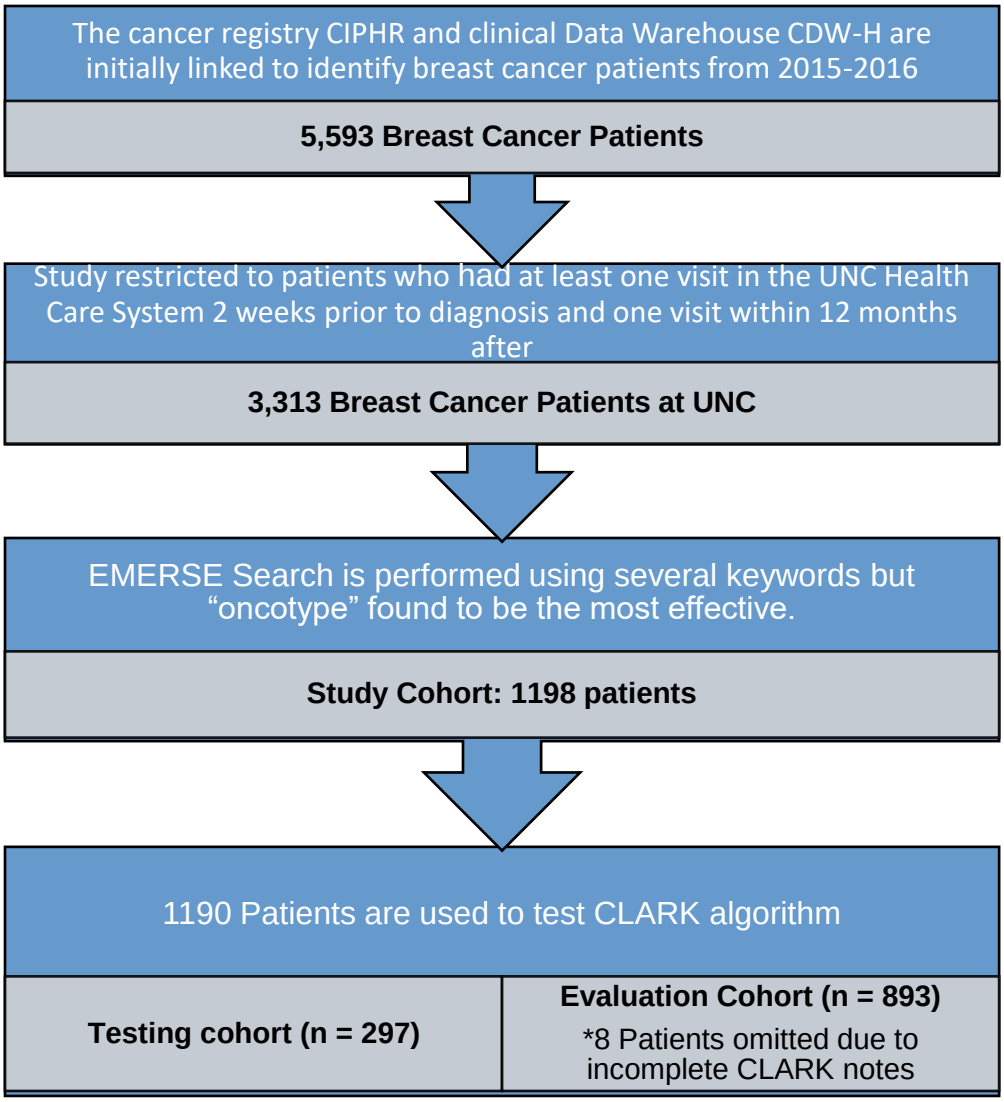

Figure 1. Population Definition

2.3: Creation of Testing Cohort to be utilized for CLARK:

As there are no other ways pulling information, manual abstraction was considered our control group (gold standard) to which the accuracy and utility of CLARK could be tested. This involved manually searching the clinical notes of each MRN identified by our "oncotype" EMERSE search in EPIC. All 1198 MRNs were manually abstracted by December 2020. The information that was abstracted included date of diagnosis, oncotype score (0-100). Each MRN was labeled based on their score: 0-17 was labeled 'low risk', 18-31 was labeled 'intermediate risk', 32-100 was labeled 'high risk', and 'NOSCORE' was assigned to patients who had no history of ODX testing results reported whether the patient or provider opted out of providing the test. Additional information abstracted was date ODX results were first reported, and a 'Yes/No' if genomic lab

report PDF was also uploaded. The 1198 patients were further split into two groups: One would be the testing cohort (or testing corpora as defined by CLARK) and the other group would be known as the evaluation cohort. The testing cohort would be used to create a “gold standard” algorithm for utilization by CLARK. The results of testing CLARK on the evaluation would be used as a comparator to our control group, manual abstraction. Each extracted MRN was then pre-labeled based on their ODX score.

2.4 Using CLARK:

The Clinical Annotation Research Kit or CLARK is a machine-learning classification software created by NCTraCS and CoVAR Applied Technologies in Durham, NC that enables “computable phenotyping in unstructured data” of free-text clinical notes.[7,8] Machine-learning software makes use of algorithms that can improve and adjust on their own. CLARK makes use of regular expressions or regex, a sequence or string of characters with additional special characters that identify a search pattern for groups of interest through literature review. CLARK uses the python “flavor” for regular expressions.[8] These regular expressions are meant to identify features that are able to distinguish one group from another.

Free-text clinical notes associated with MRNs are uploaded to CLARK in a set known as the corpus. Only free-text clinical notes within the pre-specified range of two weeks prior to date of diagnosis and 6 months after diagnosis were pulled for each MRN. It is important to note that clinical notes need to be converted to the useable format by CLARK (.json). In order for CLARK to classify properly, it needs two sets of corpora: the first is the training corpora and the evaluation corpora. The training corpora contains MRNs pre-labeled for the desired criteria. In this study, the training corpora would already have each MRN pre-labeled as low, intermediate, high or no score correctly as identified by the manual abstraction. CLARK utilizes algorithms containing regex to learn how to properly search and classify on its own by comparing the algorithm results to what the true labels are. The algorithm that ends up being utilized on the evaluation set is known as the “gold standard.” The evaluation set is unlabeled so labeling of these MRN’s will rely solely on what CLARK has learned through use of the gold standard algorithm. The goal of CLARK was for it to be able to read through clinical notes associated with MRNs and then classify or label each MRN as ‘low’, ‘intermediate’, ‘high’ or ‘NOSCORE’ based on the Oncotype DX results it identifies in the clinic notes.

2.5 Creation of “Gold Standard” Corpora and Evaluation Set:

All created regular expressions are stored in the regular expressions library. CLARK allows select regular expressions to be actively used in its algorithm (this is defined as “active regular expressions”). CLARK has options for different algorithm formats and choices on the classifier and evaluation methods. The performable algorithms formats involve: Linear SVM, Gaussian Naïve Bayes, Decision Tree and Random Forest. The evaluation method during the training process was “cross-validation” and the method chosen could be either “random” or “stratified”. Finally, the number of folds, k, could be chosen. For the purpose creating an algorithm in this project, there was no specified criteria for which algorithm format, classifier or number of folds to use; Thus, during the process of testing different combinations of options were used and compared to each other find the algorithm with the best results to be used on the evaluation set.

It was permissible to tweak the regular expressions and perform as many runs of the algorithm on the training corpus until satisfying results were achieved. It is only the evaluation corpus that would be compared to the control group. Additionally, it was important to improve the algorithm as best as possible to ensure the gold standard was as accurate as possible. In order to identify what adjustments to the algorithm were required, CLARK contains an explore page, that is only accessible after running the algorithm. This explore page reports results on the classification process including percentage of confidence CLARK was when labeling an MRN, a visual of correctly and incorrectly labeled MRNs. It contains a feature that flags keywords in individual clinic notes to see which regular expression CLARK applied to make its labeling decision; This is shown as highlighted text on the explore page within the CLARK platform. This allowed for the determination of whether or not the regular expressions written were correctly capturing the right phrases or if certain regular expressions were even impactful.

During the manual abstraction, it was noted that most ODX reporting often included either a number, 0-100, and/or was described as being “low”, “intermediate”, or high. There were variations of how this information could reported. For example, some notes might just say “Oncotype 10” or “Oncotype DX score came back yesterday as 10.” Another example might just include the phrase low such as “Patient was reported with low oncotype…”. This variability presented with the largest obstacle to writing a regular expression. Regular expressions need to be broad enough to give CLARK room to learn how to adapt to variations in phrases but also strict enough to understand how to classify phrases properly. It is insufficient to write a regular phrase that looks for the number 10 or word “low” in text as this is too broad and CLARK would

classify any note incorrectly. An example of this would be “Oncotype DX came back as 31 on 10/10/15.” CLARK would see the “10” in the date and would become confused as to whether to use “31” or “10”.

To solve this problem, a regular expression was created that required certain criteria to be met: 1) It needed to some variation of “Oncotype DX” (written as expression that contains any of those characters to then include “ODX”, “Oncotype”, etc.) 2) It needed to include a number ranging from 0-100, or the phrase “low”, “intermediate”, “or high” within at least 10 words from when “oncotype” or “ODX” was mentioned 3) If a number was identified, this number could not be included in a date (followed by a backslash or hyphen). The regular expressions were written in such a way to prevent from capturing such number. This would allow for the capture of a phase that said “Oncotype DX came back as 31 on 10/10/15” and consider the value 31 but exclude the date. This 31 would come back labeled as intermediate and not low.

2.6 Running the Evaluation Set:

Once a satisfactory “gold standard” algorithm was set. The algorithm was applied to the evaluation set. The MRNs were contained in its own .json file and uploaded to CLARK. The option for Evaluation Set was chosen in CLARK to ensure that any labels were removed from the .json file and allow CLARK to perform classification on its own.

2.7 Analysis:

Analysis in this manuscript will focus on the efficiency of CLARK. To evaluate the accuracy of CLARK, the evaluation set was compared to the labels provided by manual abstraction. Accuracy was reported as a % agreement. For the linkage portion of this project, agreement and completeness of multiple data fields across the datasets will be evaluated. Agreement will be based on weighted kappa statistics and completeness will be reported as % of data missing when the three datasets are combined. Missing data will include other information not pertaining to ODX data such as chemotherapy, endocrine therapy, types of surgery performed, etc. The exact parameters to be used will be finalized at a later date.

3. Results

All 1198 MRNs identified through EMERSE were manually abstracted; Of these 1198 MRNs, 297 were randomly selected and were used in the testing cohort for the purposes of creating the gold standard algorithm in CLARK. The remaining MRNs were used to create the evaluation

cohort (n = 901). Eight MRN's were later omitted from evaluation because they did had incomplete notes pulled within 6 months of date of diagnosis even though the EMERSION search originally identified them as a candidate for ODX testing. This was likely because ODX testing or mentioning of it did not occur until after the 6 month period of notes that were pulled for CLARK.

	True Label[a]	CLARK Label[b]	Number Correct[c]	CLARK Accuracy
NOSCORE	288	388	262/288	91.0 %
Low	347	296	271/347	78.1 %
Intermediate	208	201	179/208	86.1%
High	50	48	32/50	64.0 %
Total	893	893	744/893	83.3 %

Table 1. Distribution of Oncotype Test Results in Evaluation Set

[a]True label is categorization of ODX data based on the manually abstracted recurrence Score.

[b]Clark Label refers to number of each label that CLARK identified based on its algorithm.

[c]Number Correct returns number of CLARK-assigned labels that matched the True label

With the creation of the Gold Standard algorithm, the evaluation set was run in CLARK. CLARK labeled MRNs with 83.3% accuracy when compared to manual abstraction.

4. Discussion

The impact of precision medicine has been limited by the lack of key elements such as data extraction and linkage of multiple data sets that allows for sufficiently powered databases to examine the impact on quality of care in areas such as cancer delivery. There has been a lack of treatment and outcomes data necessary to fully evaluate the impact of precision medicine tools on patient health outcomes. This project is innovative because it attempts to correct these gaps in knowledge by linking large databases (CIPHR and CDW-H) with unstructured data found in EPIC, to obtain one data set complete with necessary clinical and claims data to study breast cancer-care delivery. The end goal of this portion of the multicomponent project has been to test data extraction methods, namely CLARK, to efficiently and quickly format genomic test results (Oncotype DX results) into structured fields so that linkage of clinical data from EPIC

back to CIPHR and CDW-H is a much easier process. As previously mentioned, this portion of the project was the first to test the utility of machine-learning software such as CLARK compared to manual abstraction in formatting large quantities of unstructured Oncotype DX data.

Overall, CLARK appeared to be quite accurate with an 83.3% agreement between the gold standard and the evaluation set. This is comparable to other studies with CLARK reporting predictive values greater than 80%.[7] The use of CLARK should prove to be a more efficient than manual abstraction due to its algorithmic approach. The time it took for CLARK to run and label the evaluation set was only a matter of minutes. This time could vary based on the RAM and processing capabilities of the computer running CLARK but this was still much faster than the several hours it took to manually abstract the ODX testing results in EPIC. Additionally, the use of CLARK it is still not a fully automatic process and requires the input of effective regular expressions to function. Researchers have to identify the keywords and phrases for CLARK to search in EPIC and then sort the data. The most time-consuming process of using CLARK was creating the gold standard algorithm. This also involved writing, adjusting and optimizing the algorithm by running the testing corpus multiple times. The limitations to the creation of the algorithm mainly lie with the ability and knowledge of the user in regard to writing regular expressions that satisfy the desired parameters for labeling. However, once a gold standard algorithm is created, labeling large sets of MRNs is a quick process. When applying this method to other genomic tests and results, the algorithm used for Oncotype DX testing will not translate exactly and new keywords and phrases will need to be identified. However, the expected outcomes from this project will help us begin to understand how to better utilize machine-learning software to be extract unstructured data as efficiently as possible.

While algorithms could be made fairly broad to capture many types of phrases, it is not possible to capture all of them without making a very strict algorithm. Additionally, CLARK is still in its first iteration and there were instances of misclassified MRNs despite the appropriate regular expression being utilized. These cases were very few. Lastly, there was also a variability in formatting for clinic notes. In some notes, the ODX note was reported in an amended surgical pathology note which often had different formatting than other clinic notes. CLARK had trouble applying algorithms to pathology notes, however, there were also few cases of these. These were not excluded from the results as they were deemed to be so small in number that it would be insignificant.

5. Conclusion

The use of machine-language learning software such as CLARK provides a useful tool for efficiently formatting unstructured data such as ODX into structured fields. While there exists some inaccuracies with CLARK, the time-efficiency cannot be overlooked when compared to manual abstraction especially when cohorts involved thousands of patients. Additionally, with continued improvements in software there will continue to be improvements accuracy and efficiency.

In the future, UNC Health may implement genomic testing modules into EPIC. This would bypass the need for some prospective data extraction methods. However, to assess retrospective data, the outcomes of this project will remain relevant and important for the assessment of precision medicine. Furthermore, the genomic module will not include every test that exists, resulting in the continued storage of unstructured precision data elements in the EHR. This approach in this study will be able to capture both historic data and data from genomic testing not included in the modules moving forward. Until genetic modules are instituted in EPIC, this project hopes to provide deeper understanding how to better abstract genomic data at UNC Health. This would be a big step in closing the disparity between methodology of precision medicine and the analysis of its results, allowing for a better understanding of cancer and cancer-care delivery. This project can also be step towards implementing similar methods in other disease states and even working outside of UNC Health.

References

1. What is precision medicine? - Genetics Home Reference - NIH [Internet]. U.S. National Library of Medicine. National Institutes of Health; [cited 2020Apr22]. Available from: https://ghr.nlm.nih.gov/primer/precisionmedicine/definition

2. Prosperi M, Min JS, Bian J, Modave F. Big data hurdles in precision medicine and precision public health. BMC Medical Informatics and Decision Making. 2018;18(1).

3. Hulsen T, Jamuar SS, Moody AR. Karnes JH. Varga O. Hedentsted S. Spreafico R. Hafler DA. McKinney EF. From Big Data to Precision Medicine. Frontiers in Medicine. 2019; 6(34):

4. Chambers DA, Feero Gregory W. Khoury MJ. Convergence of Implementation Science, Precision Medicine, and the Learning Health Care System: A New Model for Biomedical Research. JAMA. 2016; 315(18): 1941-1942.

5. Siow ZR, Boer RD, Lindeman G, Mann GB. Spotlight on the utility of the Oncotype DX® breast cancer assay. International Journal of Womens Health. 2018;Volume 10:89–100.

6. Oncotype DX: Genomic Test to Inform Breast Cancer Treatment [Internet]. Breastcancer.org. 2019 [cited 2020Apr22]. Available from: https://www.breastcancer.org/symptoms/testing/types/oncotype_dx

7. Pfaff ER, Crosskey M, Morton K, Krishnamurthy A Clinical Annotation Research Kit (CLARK): Computable Phenotyping Using Machine Learning. JMIR Med Inform 2020;8(1):e16042

8. CLARK! User and Technical Documentation [Internet]. Tracs.unc.edu. 2021 [cited 2021Apr01]. Available from: https://tracs.unc.edu/index.php/sharehub/category/2-informatics

9. Welcome to CIPHR!" *CIPHR*, 2015. Available from: https://ciphr.unc.edu/.

10. *Carolina Data Warehouse for Health*, 2008, https://tracs.unc.edu/index.php/services/informatics-and-data-science/cdw-h.

11. Hanauer, David A., et al. “Electronic Medical Record Search Engine (Emerse): An Information Retrieval Tool for Supporting Cancer Research.” JCO Clinical Cancer Informatics. 2020;4(1): 454–463

Report Addendum

Acknowledgements

The authors gratefully acknowledge the resources provided by NCTRaCS and their contributions in teaching on the use of CLARK.

Funding Support

This project was funded [in part] by a grant from the University of North Carolina Lineberger Comprehensive Cancer Center

Conflicts of Interest

There are no conflict of interests to disclose.