Progress Presentation [08/07/20]

Data Extraction Team

Cameron Breze, Will Haberkorn, Christopher Liu, George Ma, Alem Shaimardanov, David Shaw, Siddhanth Shetty, Zihan Zhang, Deniz Vurmaz

Presentation Outline

1. Introduction, Purpose, & Project Scope [Cameron & Deniz]
2. Current Project Status & Overview of Responsibilities [Cameron]
3. Data Extraction Team: Individual & Pair Contributions
    a. **Will** - BioRXIV & others (Slide 12)
    b. **Zihan** - ScienceDirect (Slide 25)
    c. **David** - Pubmed (Slide 30)
    d. **Chris** - Wiley (Slide 38)
    e. **Siddhanth** - Web of Science (Slide 42)
    f. **George & Alem** - PDFs & CV (Slide 50)
    g. **Will & Zihan** - COVID-19 update (Slide 55)
4. Machine Learning Team: Individual Contributions
    a. Justin Mae (Acute Kidney Injury project with 6 novel biomarkers)
    b. Aditya Ashtekar (Trauma Project with 2 novel biomarkers)
    c. Devashish Khulbe (Trauma Project with no biomarkers)

PROJECT SCOPE:

Knowledge Gap

1. How can rapid point of care diagnostics tests impact trauma patients?
2. Can novel biomarker panels be identified that improve trauma diagnosis?
3. How can these biomarkers be best measured at the point of care?
4. What multivariate models are most effective in trauma diagnosis?
5. Can other data fields be fused with trauma biomarker tests to create more effective models.

Innovation

This approach leverages the synergies among:

❑ Analyte capture and interrogation technologies High-Content Analysis (HCA) methods

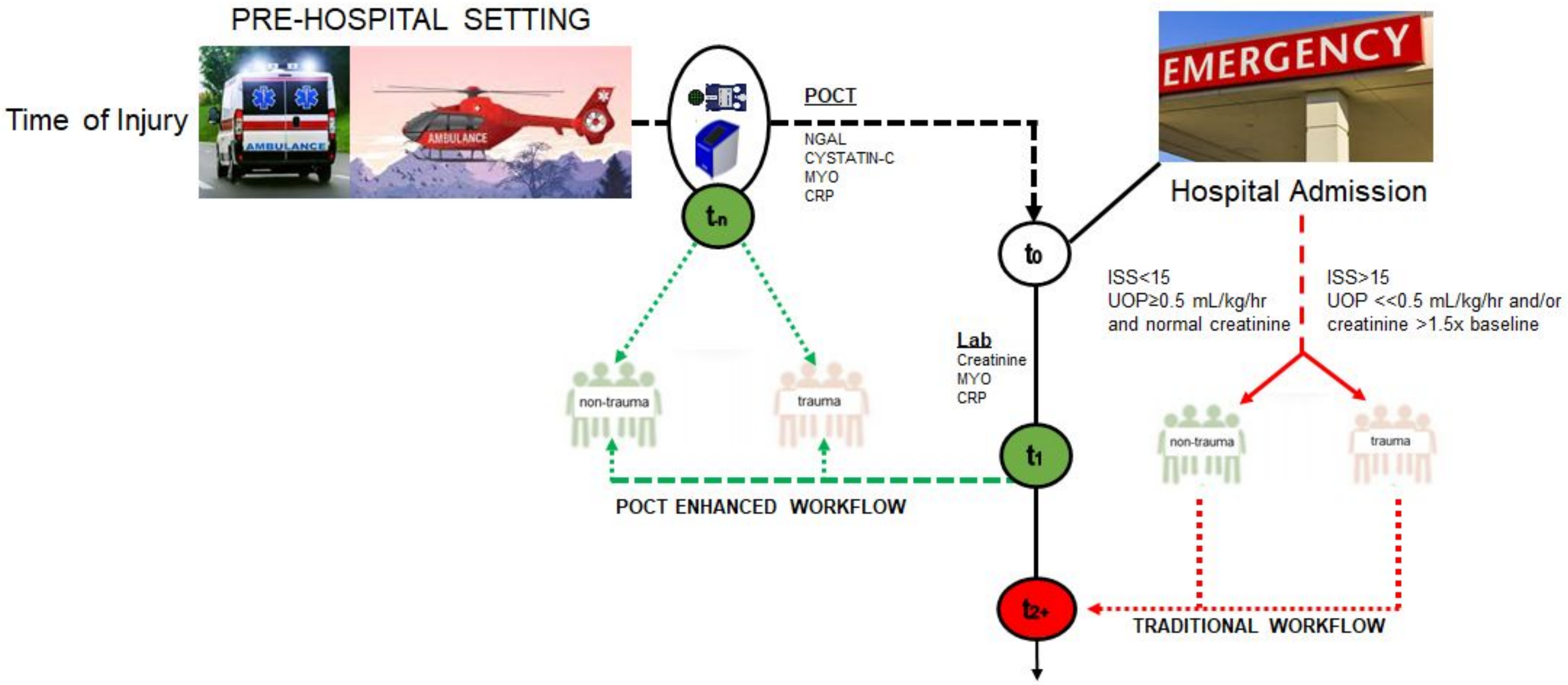
❑ Machine learning algorithms

...and has resulted in...

1. Carefully developed and validated predictive models that helps the performance of diagnosis by expert clinicians/surgeons
2. Largest database of protein-based biomarkers for prospectively recruited Trauma cases. NLP embedding algorithms can be trained for diagnostic performance characteristics in trauma related articles.
3. New insights into the organ specific injury, hemorrhage and severity of trauma differences across the trauma spectrum

OBJECTIVE: Develop robust classification models from trauma-on-a-chip measurements that alternate diagnostic performance of gold standard approach

Trauma Diagnosis WorkFlow



PRE-HOSPITAL SETTING

Time of Injury

t_{-n}

POCT

NGAL
CYSTATIN-C
MYO
CRP

t_0

Hospital Admission

ISS<15
UOP≥0.5 mL/kg/hr
and normal creatinine

ISS>15
UOP <<0.5 mL/kg/hr and/or
creatinine >1.5x baseline

Lab
Creatinine
MYO
CRP

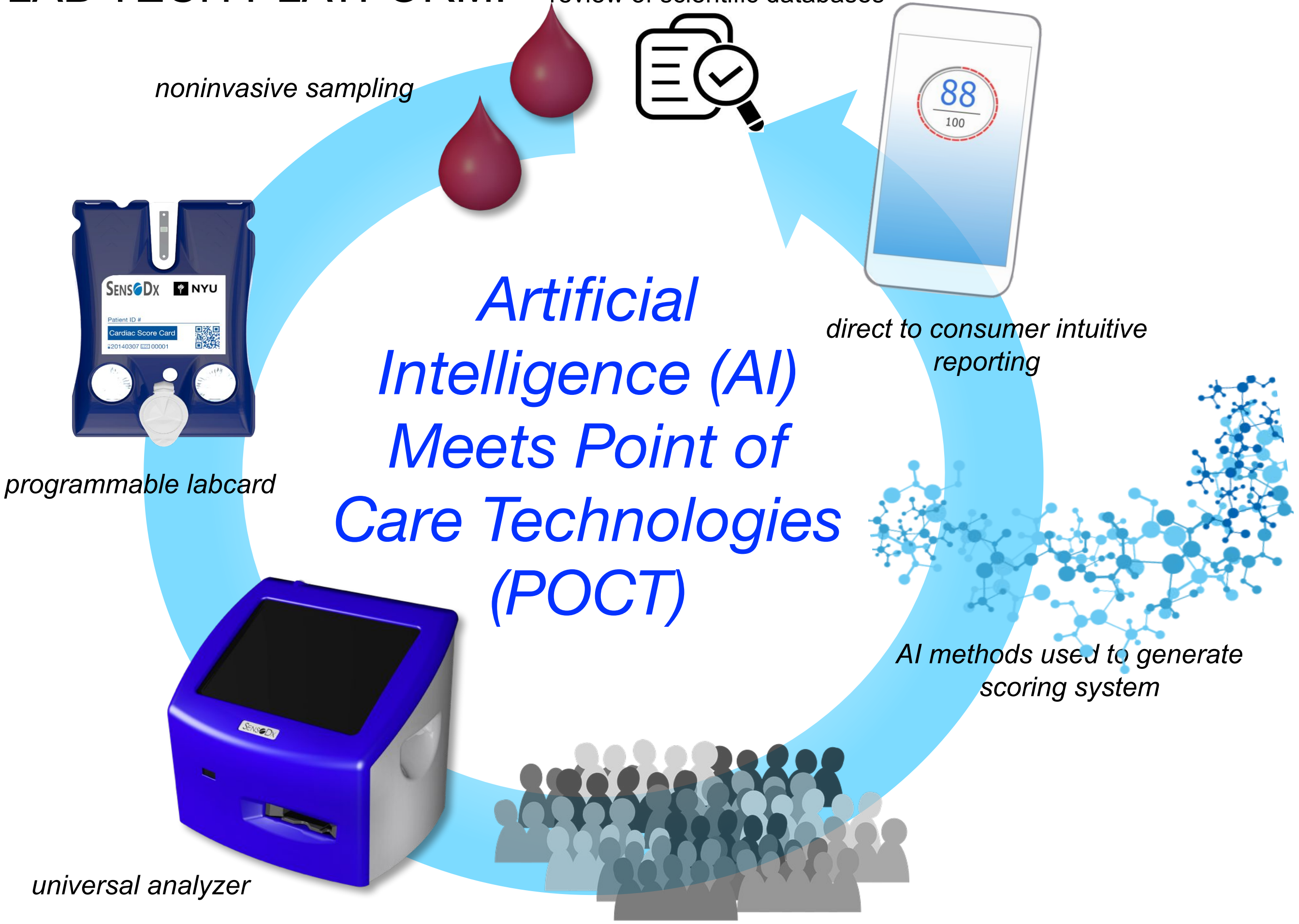
t_1

POCT ENHANCED WORKFLOW

t_{2+}

TRADITIONAL WORKFLOW

MCDEVITT LAB TECH PLATFORM:



noninvasive sampling

Systematical meta-analysis review of scientific databases

direct to consumer intuitive reporting

AI methods used to generate scoring system

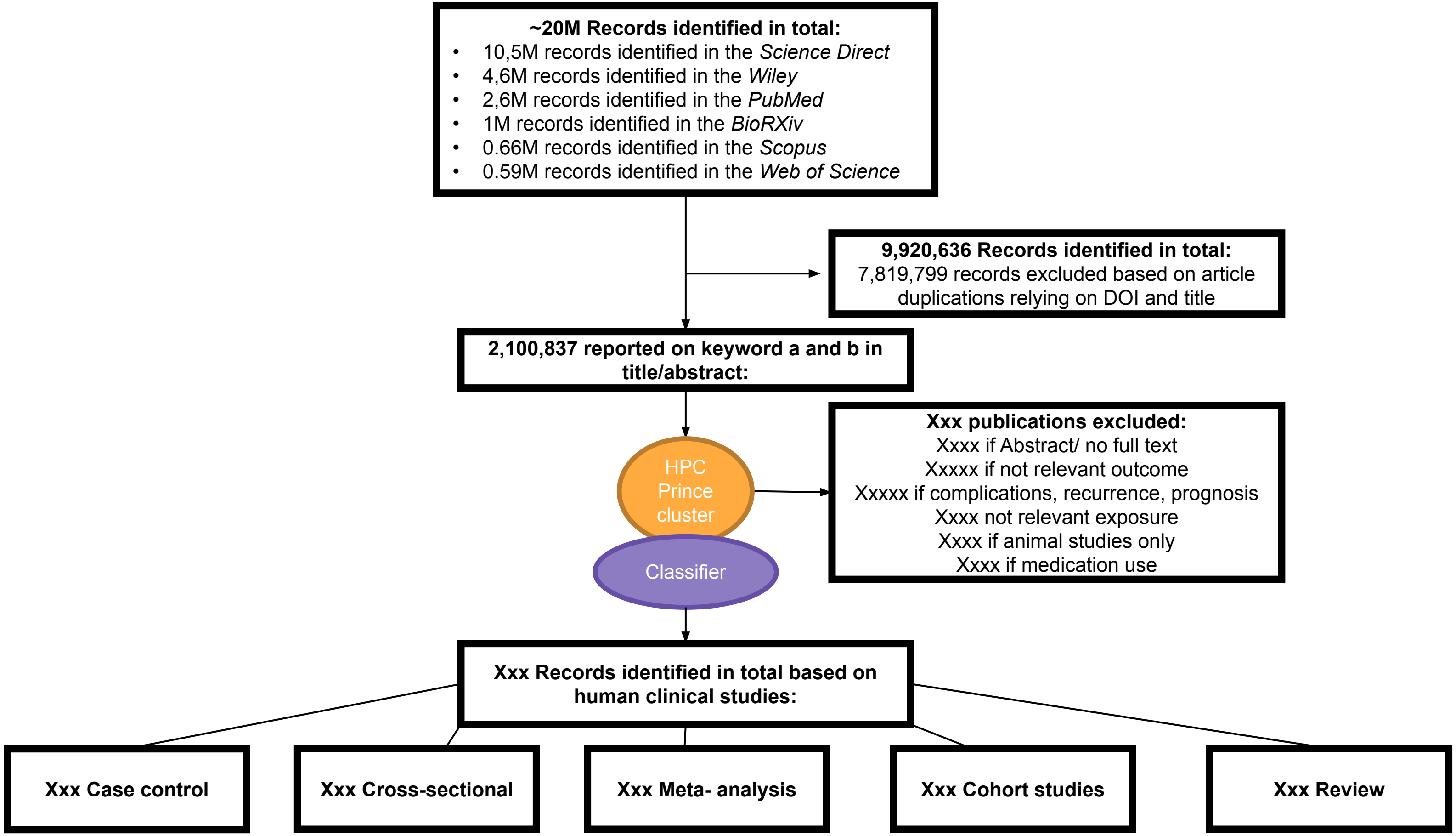
well designed trials used to establish training methods

universal analyzer

programmable labcard

Artificial Intelligence (AI) Meets Point of Care Technologies (POCT)

PROJECT FLOWCHART



~20M Records identified in total:

- 10,5M records identified in the *Science Direct*
- 4,6M records identified in the *Wiley*
- 2,6M records identified in the *PubMed*
- 1M records identified in the *BioRXiv*
- 0.66M records identified in the *Scopus*
- 0.59M records identified in the *Web of Science*

9,920,636 Records identified in total:
7,819,799 records excluded based on article duplications relying on DOI and title

2,100,837 reported on keyword a and b in title/abstract:

Xxx publications excluded:
Xxxx if Abstract/ no full text
Xxxxx if not relevant outcome
Xxxxx if complications, recurrence, prognosis
Xxxx not relevant exposure
Xxxx if animal studies only
Xxxx if medication use

Xxx Records identified in total based on human clinical studies:

Xxx Case control

Xxx Cross-sectional

Xxx Meta- analysis

Xxx Cohort studies

Xxx Review

NLP & ML Approach Of Integrated Multi-Parameter Trauma Diagnostic Tool

NLP PROJECT GOAL:

We use a systematic literature search to gain access to an understanding of things like:

1) which biomarkers work best for the various types of trauma

Tool: Text mining by Python

2) we bridge between existing published studies to develop a multivariate model that incorporates both biomarker data and other fields.

This second area would involve machine learning for prediction model for diagnostic performance.

Tool: Data mining by Computer vision, CNN, Monte Carlo Simulation

ML PROJECT GOAL:

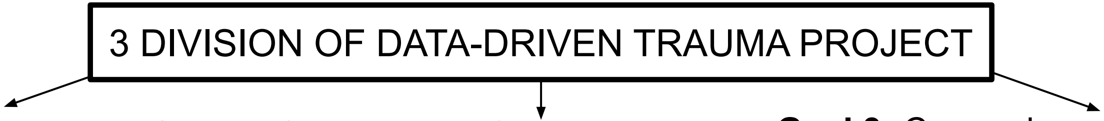
We use a different ML models and combination of them to gain access to an understanding of things like:

1) How can we improve diagnostic performance of current trauma models
2) What's the most important features for multivariate trauma diagnostic model?
3) Novel biomarkers can diverse and increase the model performance/accuracy?

Tool: Lasso, LR, FR, SVM, DNN, Super Learner

OBJECTIVE: Develop robust classification models from trauma-on-a-chip measurements that alternate diagnostic performance of gold standard approach

TRAUMA PROJECT MODEL DEVELOPMENT AND APPROACH:



3 DIVISION OF DATA-DRIVEN TRAUMA PROJECT

Goal 1: Selection of biomarkers

Goal 2: Comparison of current conventional markers and scoring system

Goal 3: Comparison of model development with novel and conventional biomarkers and time-course study

Systemical Review- Meta Data Analysis

- The dataset is extracted from 12M publications.
- High Performance Computing Platform
- Collected papers based on Different type of clinical trials, study design
- Novel biomarkers: MYO, PCT, D-Dimer, CRP, FABP-1, FABP-2, Protein-C, N-GAL, Cystatin-C, Properdin, C5a, HMGB-1 are available.

Outcome: prediction of the most important biomarkers for organ specific trauma and poly-trauma, comparison of multivariate LR model for conventional vs novel biomarkers

National Trauma DataBank

- Conventional markers and variables (not includes biomarker)
- **60,000** trauma patient dataset (year 2016)

- **Outcome:** probability of trauma mortality/ severity of trauma
- **Models:** LR, Lasso, Forest Random, DNN
- Performance comparison, precision (reproducibility), AUC comparison
- Trained dataset to create scoring system: 60,000
- Validation dataset: 60,000 (year 2015)

Trauma (from a research paper: raw dataset)

- The data set included 2,397 variables on **1,494** patients.
- Activation of Coagulation and Inflammation in Trauma study (ACIT)
- Single-center prospective cohort study
- Time-course study is available
- D-Dimer, Protein-C, coagulation factor protein-biomarkers are available

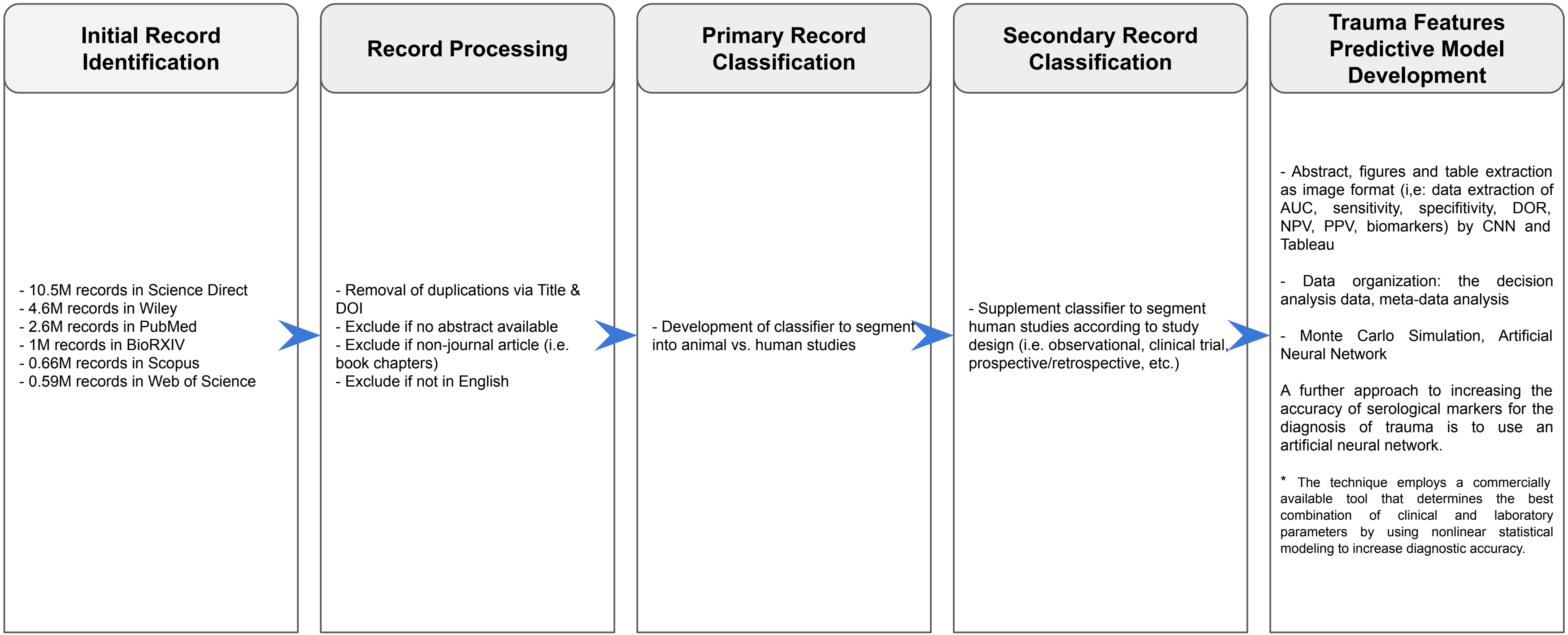
Outcome: prediction of multiple organ failure/ severity of trauma

Models: LR, Forest Random, Lasso, DPP

- Selection of the most important features, Performance comparison, precision (reproducibility), AUC comparison
- Validation dataset: 60 Patient samples McDevitt Lab-Bellevue Hospital

Next steps

Project Flowchart



Initial Record Identification

- 10.5M records in Science Direct
- 4.6M records in Wiley
- 2.6M records in PubMed
- 1M records in BioRXIV
- 0.66M records in Scopus
- 0.59M records in Web of Science

Record Processing

- Removal of duplications via Title & DOI
- Exclude if no abstract available
- Exclude if non-journal article (i.e. book chapters)
- Exclude if not in English

Primary Record Classification

- Development of classifier to segment into animal vs. human studies

Secondary Record Classification

- Supplement classifier to segment human studies according to study design (i.e. observational, clinical trial, prospective/retrospective, etc.)

Trauma Features Predictive Model Development

- Abstract, figures and table extraction as image format (i,e: data extraction of AUC, sensitivity, specifitivity, DOR, NPV, PPV, biomarkers) by CNN and Tableau

- Data organization: the decision analysis data, meta-data analysis

- Monte Carlo Simulation, Artificial Neural Network

A further approach to increasing the accuracy of serological markers for the diagnosis of trauma is to use an artificial neural network.

* The technique employs a commercially available tool that determines the best combination of clinical and laboratory parameters by using nonlinear statistical modeling to increase diagnostic accuracy.

Key Results: Web Scraping

Description		Status	Area(s) of concern	Next Steps
Result 1	Generate preliminary database of journal articles relevant to trauma keywords and store as SQLite DB with information on DOI, URL, article type, etc.	Completed.	Heterogeneity in DB columns being populated. Some database sources provide more information than others.	---
Result 2	First pass at excluding journal articles. Remove articles not in English, incorrect article type, duplicates.	Completed.	Each database has a subset of possible article types (or no article distinction).	Establish an “in” and “out” list for each specific database for article type
Result 3	Bulk download of applicable PDFs and storage of text files in HPC cluster for further analysis.	In Progress; waiting on WoS	Copyright issues associated with bulk downloads have come to light with Pubmed. These will also need to be checked in other databases.	Follow up with each database with project proposal to gain permission to bulk download or set up a private API to download.
Result 4	Second pass at excluding articles. Sort articles by animal studies vs. human studies.	Completed.	Unclear whether exclusion of animal studies or inclusion of human studies will be more effective	Annotate a group of PDFs and write 2 scripts (one for each strategy) and compare results to determine which is more accurate
Result 5	Re-create heat map analysis. Compare to initial heat map results for trauma insights.	Not Started.	None as of yet.	---

Project Timeline

05-15-2020	Last work progress meeting, setup HPC access to “PDF” folder
05-22-2020	Resolve copyright issues with SciHub, setup Mendeleev for open source
05-29-2020	Setup scratch folder access, begin PDF to text scripts, install selenium on HPC
06-05-2020	CV project initial research, present biomedical ML/AI papers, removal if no full text avail.
06-12-2020	Strategy realignment, CV continues, language detection, bulk download policies
06-19-2020	Initial results from CV arm, download policies, setup access Brooklyn HPC, Justin (ML)
06-26-2020	Test individual meetings, Django site, initial classifier research and strategy
07-03-2020	Individual meetings, kernel speed in CV, first classifiers built and results compared
07-10-2020	Annotated AKI papers, COVID project overlaps, classifier data joined together
07-17-2020	Second classification round started, waiting on WoS script, NLP overlaps with COVID/CV
07-24-2020	False positives with NLP CV, COVID data sharing agreements, split WoS scripts
07-31-2020	Plan to recreate heat map analysis, continue WoS downloads, combine classifiers into 1

Team Responsibilities

Name	What You Worked On	Next Steps
Cameron Breze	Managed data extraction team, led weekly meetings	
William Haberkorn	BioRxiv -Change search strategy to get better results, obtain the relevant articles and downloads Animal vs Human -Collect, label, process articles for the dataset,build an initial TF-IDF based model, find more optimal feature subsets, finalize model Clinical Study Design-Initial data collection and labeling Covid- Initial data search, begin downloads	● Find and remove non trauma related/studies not in English from BioRxiv articles. ● Implement/improve the Animal classifier ● Build the Clinical Study Design classifier and its dataset. ● Complete downloads for COVID articles and cluster.
Christopher Liu	Wiley - Downloaded articles as PDFs from database Parsed PDFs into text classify articles into animal vs human through logistic regression classifier w/ Will Labeling of data by clinical study design	● Apply the same steps used in the animal vs human classifier and extend it to different types of clinical studies.
George Ma	Scrape Scopus abstracts and parse text from PDFs	
Alem Shaimardanov	● Systematical Review and Meta-data Analysis, Monte Carlo Simulation ● Collected information about abstracts, journal types, keywords, figures and tables of 20 selected studies (selected 1 Nature paper: systematical review of Cystatin-C)	● Automate extraction of detailed assessment parameters (AUC, sensitivity, specificity) from the selected studies
David Shaw	Automated PubMed article extraction and labeling	
Siddhanth Shetty	Downloading Web Of Science Articles and filtering english articles	
Zihan Zhang	ScienceDirect - Automatically download relevant pdfs based on given keywords. Construct sqlite database. Convert pdfs to texts and filter for Animal vs. Human articles. Covid - Initial Data search, finish downloads.	Use NLP to analyze Covid articles. Cluster and find similarity between texts to conclude certain relevant keywords.
Deniz Vurmaz	Defined the project roadmap and helped strategizing the project, helped team when they need support for Trauma, Covid-19 project, led ML team, provided organized dataset.	● Helping to build Monte Carlo Simulation, Model Development

Individual Efforts

- Will - BioRXIV & others ◣
- Zihan - ScienceDirect ◣
- David - PubMed ◣
- Chris - Wiley ◣
- Siddhanth - Web of Science ◣
- George & Alem - PDFs & CV ◣
- Will & Zihan - COVID-19 update ◣

BioRxiv and Article Classification

William Haberkorn

Changing Approach for BioRxiv

Issues: The BioRxiv internal search method was poor, and would provide primarily irrelevant results when searching with keyword pairs, and excluding the search to “match phrase” and “title or abstract” still resulted in many unwanted articles, and almost all were repeated many times after going through the keyword combinations.

BioRxiv also has no API that allows for download of articles based on search terms and filtering articles so everything had to be done through a web scraper, and going through every keyword combination and storing information about the 1,000,000+ results of the previous search was taking too long.

Web Scraping and Downloads

To avoid the issues when searching BioRxiv with a keyword pair, I decided to store the DOI of every article that appeared in the results of keyword A and the DOI for all results related to keyword B. I then retrieved the article information only when the DOI was in both results lists, which lead to more accurate results overall (although still imperfect).

This was also much faster than BioRxiv's search method and instead of taking over a week, the script finished in under 4 hours.

Keyword B resulted in ~500,000 articles, Keyword A resulted in ~300,000 articles

The unique intersection was about 5,000 articles.

All articles were separated by whether they had been published or if they were a preprint, and slightly more than half of the articles were preprints. Some of which are still not relevant, and if not filtered out with the upcoming classifiers I can write a separate model to determine if it is trauma related.

Animal vs. Human Classifier

Idea: transform every article in the manually labeled data set (N=100) into a TF-IDF vector, reduce the dimensionality of the entire corpus to a manageable size, and build a classifier that will separate the data into the appropriate class.

Initial roadblocks:

- Initial feature extraction and preprocessing of the text greatly affects the results, and processing/storing the data can take up a lot of time/space.
- Many BioRxiv articles were not clearly in either class, or were in both, so gathering enough data was difficult.
- Reducing high dimensional data into a form that performs well in a classifier is ultimately trial and error.

Animal vs. Human Classifier - First Models

Initially ~100 documents were labeled by Christopher and I and used in the model.

1. Every document was read in and all non-words or stop words (i.e. ‘the’,’an’,’it’) were removed, and the text file was cleaned of any errors from the conversion.
2. Each document was then replaced with a .txt file containing every unique word in the article next to its frequency as long as the word appears more than twice in the article. N-grams were stored in place of their components if they occured at least 8 times.
3. While the files are read all the unique words across all documents are stored in another file, next to the number of articles they appeared in.
4. All documents are read in again and each word’s frequency is multiplied by its inverse document frequency.

$$\log(\ N / |\{\text{Document}:\text{word} \in \text{Document}\}|\ )$$

Animal vs. Human Classifier - First Models

Features whose sum TF-IDF score was less than 0.01 were removed.

The feature set still had 5,000 features, so to reduce this I used both a filtering method that utilized one-way ANOVA to get the N most relevant features. 100,150,200,250,500 were tried. But the models accuracy was stuck at ~90%, so I used PCA to transform the dataset with 500 features to one with 100 to improve the results.

With the first dataset, I tried ***Random Forest, SVM, Linear Regression, and Naive Bayes. Random Forest achieved 97% accuracy and SVM and Linear Regression had ~95% accuracy.***

Animal vs. Human Classifier - More Data

Main Issues: The first model had a sample size that was too small for a more general model, and while immediately converting the text into the articles word dictionary was faster and saved space it made changing anything in the preprocessing phase very difficult.

Solution: I built a scraper to gather pubmed articles about trauma and filtered by article type and human/animal, and gathered 100 more articles in each class. The files were converted to text and the cleaned file was stored instead of a dictionary to allow for quick modification of features and processing.

I filtered out words which only appeared in <15% or >90% of documents. This reduced the feature set from 18,000 to 780 (other percentages were tried but this seemed to be the best result).

Many of the 780 words that the word list reduces down to are not helpful to the classification. To reduce the dataset further I implemented a Sequential Backwards Selection search method to obtain a better subset of 150 words. This can only be done with a smaller reduction as it’s computationally expensive, and this took 2 days to complete on HPC. Thus, most of the feature set had to be filtered out prior.

I found this to be the subset size that produces the best results.

Animal vs. Human Classifier - Second Approach

When using the same models as before the accuracy was 83% for Random Forest and SVM.

To reduce the error and noise of the model I decided to use an ensemble approach with multiple models and boosting methods, which utilize multiple classifiers to make a prediction with an accuracy higher than any of its components. The boosting methods were incorporated to prevent the model from overfitting.

I tried many models and the the accuracy reached 95% after tuning hyperparameters, but the model was very complicated and used 5 sub machine learning algorithms in order to make a prediction. Also, many of the PDF’s did not properly convert to a text format which seemed to be making prediction more difficult so I decided to gather more data and fine-tune the preprocessing.

Animal vs. Human Classifier - Final Model

I downloaded a few hundred more articles and re-converted the PDF’s to text, and discarded files that would only convert a couple sentences or had other conversion errors and stored their filenames to be processed at a later point using Tesseract.

Christopher also downloaded a few thousand articles from Wiley. The final dataset had about 8,000 records, but only ~2,000 human articles. I only considered 2,000 animal records in the model to ensure an even class distribution (without this the Human recall was lower, but overall accuracy was 98-99%).

Using the same set of 150 features the Random Forest model had an accuracy of 98.3% with 5-Fold Cross Validation.

Recall:	Human-98.6%	Animal-98.0%
Precision:	Human-97.7%	Animal-98.8%

All other models had a similar accuracy. The model doesn’t appear to be overfitting as the testing and training accuracy are nearly the same, and I’ve limited the depth of the classifier to reduce the model’s ability to overfit.

Animal vs. Human Classifier - Issues

The primary issue with implementing this model will be dealing with damaged and difficult to read PDFs, as would be true with any text based model.

All the BioRxiv articles we are using are published, which can all be downloaded at other locations which I have found to be better for text conversion. Since there aren't very many articles from BioRxiv, I can download the cleaner formats from other websites, and when the script comes across issues from another database's article it can store it in a temporary directory to be read with Tesseract later on.

To ensure that the model is not overfitting, I can obtain more articles and verify the results.

Clinical Study Design Classifier

Converting the articles to TF-IDF matrices may work to determine the overall Study Design of the paper, but in order to classify the articles within each of these designs as {“Analytical”,”Observational”,”Descriptive”} it would likely not be enough.

To classify documents into these subcategories would require contextual information, which is something TF-IDF cannot provide even using large N-Gram models.

Word2Vec (or another word embedding) would be the best approach as it considers semantically similar words and sentences. It is also a neural network based model which can handle large feature sets without hurting performance.

Clinical Study Design Classifier

Instead of using TF-IDF to classify the overall design, and then using Word2Vec for the subclasses, I will be using Word2Vec to classify the articles with multiple labels at once.

I believe this will minimize the error by limiting the problem to one model, and many of the class sets (i.e. “Cohort Study” & “Prospective”) will be codependent.

Clinical Study Design Classifier - Approach

The first step will be to create a large dataset of tagged articles. The larger classes like “Meta-Analysis” and “Case Series” are easy to download with scrapers by searching journals with well defined filters, but terms like “Analytical” and “Descriptive” will require some manual tagging (although will sometimes be directly next to the study design term).

I have started downloading articles for Randomized Control, Cohort, Meta Analysis etc. but would not be able to label enough articles with the appropriate sub terms.

I can create an open CSV file with the link to the article and share it on Google Drive if people would be willing to help out. I also plan to create an sqlite database file to store the information feeding into the model since this project will need much more data, and I can share this and some compressed form/location of the data on the cluster if other people want to use it for their classifier.

Science Direct

Zihan Zhang

Download Update

1. Exclusion criteria for Article Types
    a. Only remain Short Communication, Review Article, Research Article, Case Report
2. Remove duplicated articles based on DOI

Outcome: Effectively reduce workloads from *7 millions to 566470*

Download Update

IP blocked after 2000 download using NYU HPC

1. Add time.sleep() function: 25-30 seconds.
    a. Pretend to be random behavior.
    b. Uneffective and waste of time. 2-days validity period of session_id.
2. Brooklyn Research Cluster
3. Multithreading
    a. Improve security and flexibility of database. Resume from breakpoint.
    b. Enhance CPU utilization.
    c. Program responding quicker.
    d. Adding TD_Controller and TD_Worker class to distribute projects from queue.
    e. Upper bound: Downloading speed

→ Combine

Download Update

Store PDF file as _cfg.type_blob format in the database. Easily manageable.

For multiple authors, use cross-join and break down them into independent keys.

Classifier: Keyword Frequency

Convert PDFs into texts and search keywords frequency.

(“Patient”, “Human”, “Children”, “People”, “Male”, “Female”, “Adult”, “Trauma Patients”, “Homo Sapiens”)

PubMED

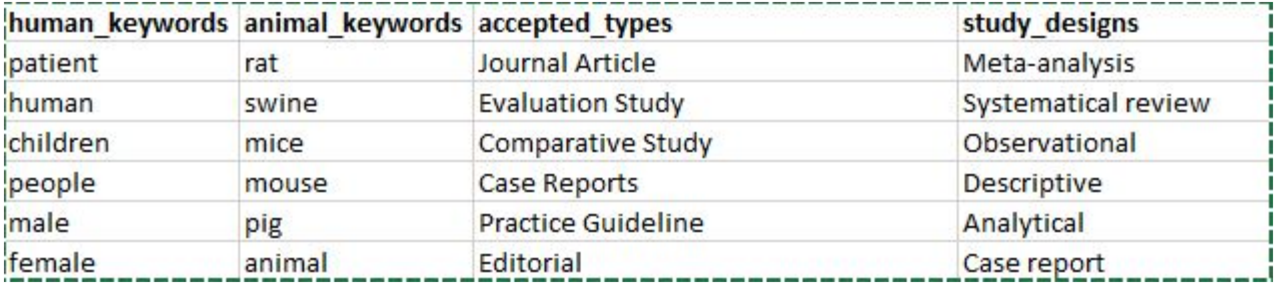
David Shaw

Corrections

1. The search function was modified to enforce keyword integrity.
   a. Number of articles found was reduced to 55,437.
2. PMID is added as a new column.

Keywords/Exclusion Criteria

- Keywords from the Word documents are aggregated and saved in Excel format.
- Comparisons ignore cases



human_keywords	animal_keywords	accepted_types	study_designs
patient	rat	Journal Article	Meta-analysis
human	swine	Evaluation Study	Systematical review
children	mice	Comparative Study	Observational
people	mouse	Case Reports	Descriptive
male	pig	Practice Guideline	Analytical
female	animal	Editorial	Case report

Downloads

Legality

PubMed limits bulk downloads to the Open Access Subset. Each article in the subset has either the commercial or noncommercial license.

Methodology

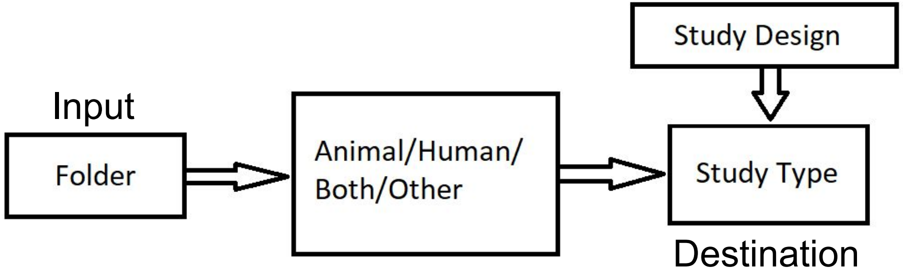
- Script downloads archives directly from PubMed FTP server.
  - Each archive contains an .xml file and a combination of .pdf for the article itself, and .jpg/.gif for tables and images in the article.
- All downloaded archives are unpacked and saved in ./PDF with PMID as their folder names.

Downloads

Results

- 12,017 out of the 55,437 articles were found and downloaded from the server.
- 297 out of the 12,017 downloaded archives each has one of the following problems.
  - No .pdf in folder.
  - .pdf file does not have the article title returned by the search function.

Sorting



For each folder in ./PDF, if it contains at least one .pdf, the .pdf file that contains the matching article title is located and scanned to determine its categories. The folder is

1. Categorized as either a human, animal, both, or other study.
2. Moved to a subfolder for its study type.
3. (Human Studies Only) Labeled with study design keywords.

I.E. "./PDF/00001/" -> "./Human Studies/Clinical Trial, Phase I/00001/"

Sorting

Exceptions

- Folders with missing .pdf or the .pdf without matching article title remain in ./PDF.

Records

./OpenAccess.csv

- General and license-related information for articles downloaded from FTP server.

./downloads_summary.csv

- Summary for downloaded archives. It contains
    - Original and sorted file paths
    - Article information (PMID, article type, title)
    - Statistics used for sorting
    - Rejection status
    - Study design labels

Wiley and Article Type Classifier

Christopher Liu

1. Downloaded PDFs

Used a script to download PDFs from Wiley

Fairly simple, didn’t run into any problems here

2. Animal vs Human classifier

First tried to separate animal vs. human studies through keyword detection, found not accurate enough

Built a logistic regression classifier to guess whether animal/human based on frequency distribution / presence of words in article.

~2000 animal studies and ~2000 human studies used as data for the classifier, tested on known articles from Wiley. Tests show >95% accuracy.

Article Type classifier

Extend classifier to determine study type (observational, analytical, retrospective, prospective, cohort, etc.)

Articles aren’t very clearly defined here

First steps mean labeling observational vs. retrospective

Extend the animal vs. human classifier

Web of Science and Language Classifier

Siddhanth Shetty

Tasks

Primary task: Web of Science PDF download

Secondary task: Classify articles by language

Articles to be downloaded

- Source: Database containing article name, DOI, publish date
- Duplicates removed
- Totally 339,539 Articles from Web Of Science
- Articles need to be divided by article type(eg. Meeting abstract, Proceedings etc.)

Roadblock

After reaching out to Clarivate we confirmed there's no API to download full-text PDF’s from WebOfScience

Workaround

Use of Kopernio extension along with Full-Text links available for some papers to download papers and separate them by article type

Drawbacks

- Able to download approximately one-third of all the articles
- Slow as it has to wait for the kopernio extension to locate the paper as well as download each file(Cameron and Will are helping in parallelizing the download)
- Cannot run on the cluster as the extension doesn’t work without a chrome instance running(cluster runs it in headless mode)

Current status

16,161 pdfs downloaded out of a potential 49,369.

Secondary Task

- Using Langdetect(python language classifier) to filter out pdfs that are not mostly in English(at least 80% in English)

Data Visualization of PDFs w/ Computer Vision

Alem Shaimardanov, George Ma

Background information

Z. Yong, X. Pei, B. Zhu, H. Yuan, W. Zhao, *Predictive value of serum cystatin C for acute kidney injury in adults: a meta-analysis of prospective cohort trials*. Scientific Reports. 2017; 7:41012 | DOI: 10.1038/srep41012

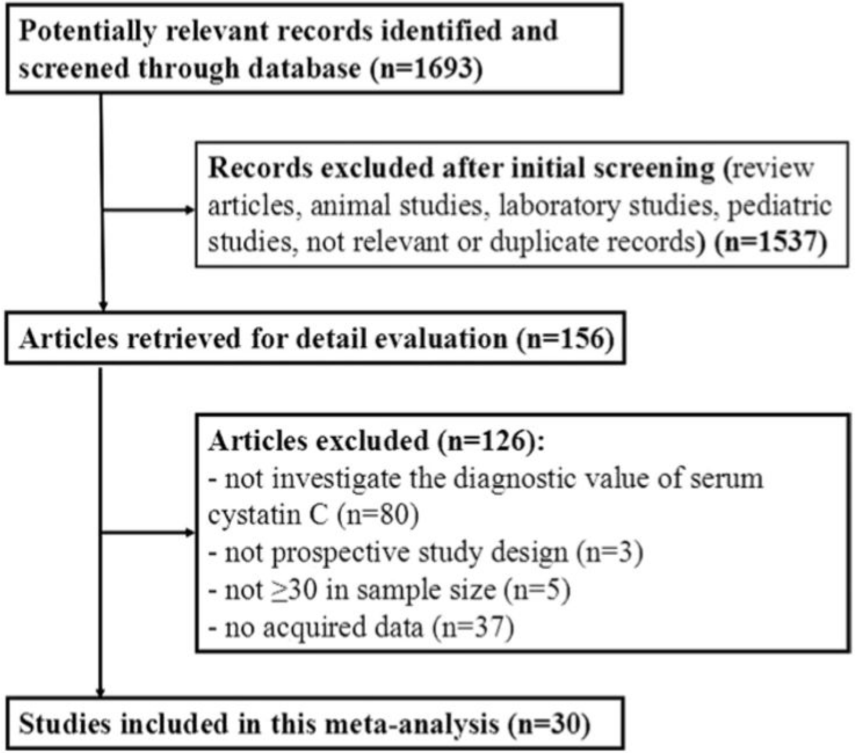
Aim: to investigate the overall diagnostic accuracy of serum cystatin (Scys) for Acute Kidney Injury (AKI) in adults, and further identify factors affecting its performance.

Studies were retrieved from PubMed, Embase, Web of Science and Cochrane Library.

Background information

Study selection: (1) prospective cohort study, (2) adults, (3) sample size ≥ 30, (4) original data of sensitivity and specificity, (5) AKI diagnostic criteria.

If any disagreement existed, two investigators would check and discuss about the full text.



Potentially relevant records identified and screened through database (n=1693)

Records excluded after initial screening (review articles, animal studies, laboratory studies, pediatric studies, not relevant or duplicate records) **(n=1537)**

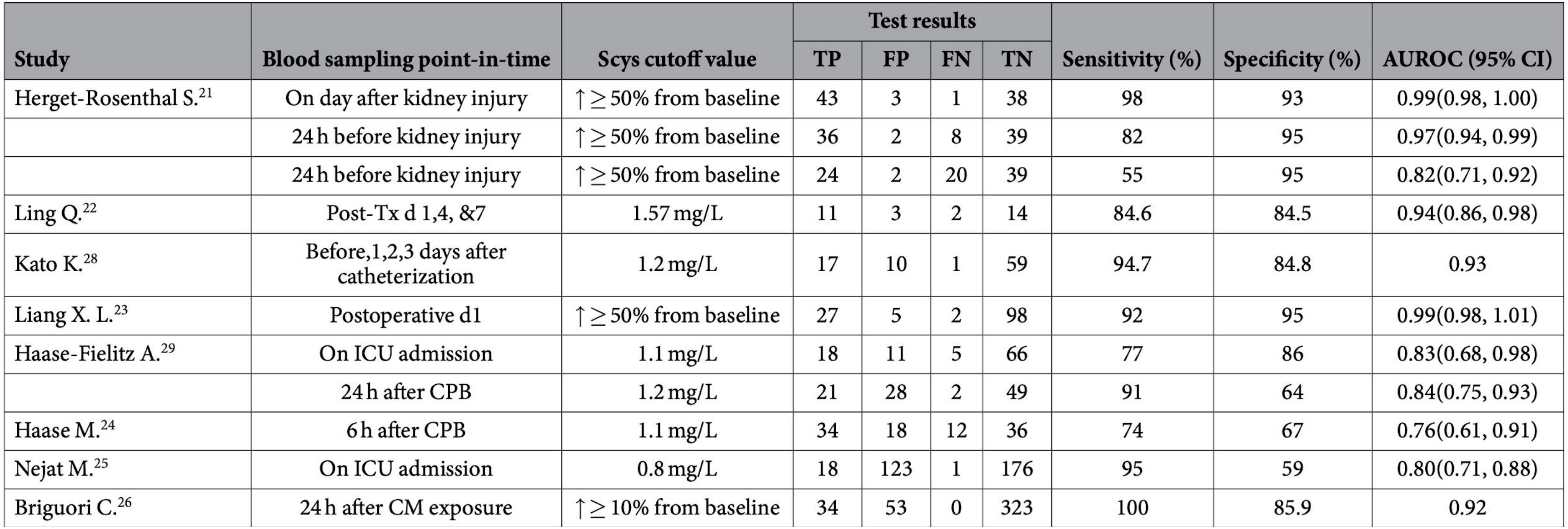
Articles retrieved for detail evaluation (n=156)

Articles excluded (n=126):
- not investigate the diagnostic value of serum cystatin C (n=80)
- not prospective study design (n=3)
- not ≥30 in sample size (n=5)
- no acquired data (n=37)

Studies included in this meta-analysis (n=30)

Figure 1. Flow chart of study selection.

Table 2. The accuracy of serum cystatin (Scys) at various blood sampling point-in-time and cut-off value.



Study	Blood sampling point-in-time	Scys cutoff value	Test results				Sensitivity (%)	Specificity (%)	AUROC (95% CI)
			TP	FP	FN	TN			
Herget-Rosenthal S.[21]	On day after kidney injury	↑≥50% from baseline	43	3	1	38	98	93	0.99(0.98, 1.00)
	24 h before kidney injury	↑≥50% from baseline	36	2	8	39	82	95	0.97(0.94, 0.99)
	24 h before kidney injury	↑≥50% from baseline	24	2	20	39	55	95	0.82(0.71, 0.92)
Ling Q.[22]	Post-Tx d 1,4, &7	1.57 mg/L	11	3	2	14	84.6	84.5	0.94(0.86, 0.98)
Kato K.[28]	Before,1,2,3 days after catheterization	1.2 mg/L	17	10	1	59	94.7	84.8	0.93
Liang X. L.[23]	Postoperative d1	↑≥50% from baseline	27	5	2	98	92	95	0.99(0.98, 1.01)
Haase-Fielitz A.[29]	On ICU admission	1.1 mg/L	18	11	5	66	77	86	0.83(0.68, 0.98)
	24 h after CPB	1.2 mg/L	21	28	2	49	91	64	0.84(0.75, 0.93)
Haase M.[24]	6 h after CPB	1.1 mg/L	34	18	12	36	74	67	0.76(0.61, 0.91)
Nejat M.[25]	On ICU admission	0.8 mg/L	18	123	1	176	95	59	0.80(0.71, 0.88)
Briguori C.[26]	24 h after CM exposure	↑≥10% from baseline	34	53	0	323	100	85.9	0.92

Our goal: Create a similar table by automating data extraction from 30 selected studies.

Problem: Some PDFs cannot be read by traditional text parsing packages (e.g. Tika, PDFminer, pyPDF)

Solution:

Computer vision based approach, first render PDF, then use OCR network to ‘read’ PDF and get text regions. (Done)

Additional step:

Filter out non-essential regions. (keyword based approach)

Separate charts from text blocks. (difficult, non-uniform)

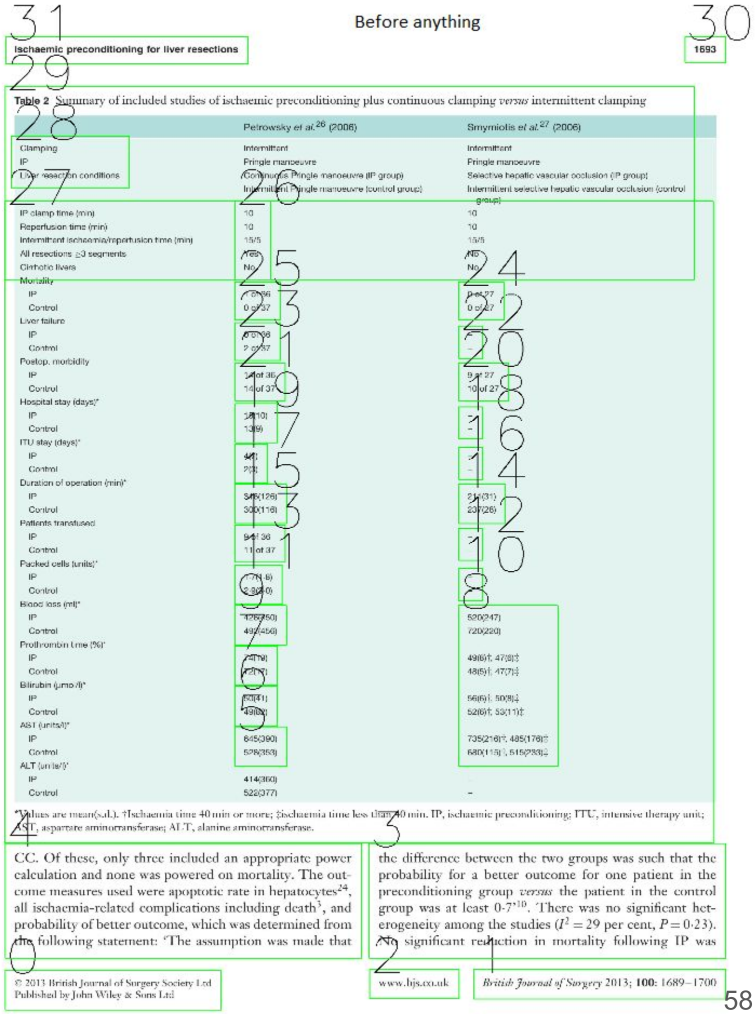
Computer Vision approach

Technology behind this:

pdf2image/poppler

opencv-python

pyTesseract



Ischaemic preconditioning for liver resections 1593

Table 2 Summary of included studies of ischaemic preconditioning plus continuous clamping *versus* intermittent clamping

	Petrowsky *et al.*[26] (2006)	Smymiolis *et al.*[27] (2006)
Clamping	Intermittent	Intermittent
IP	Pringle manoeuvre	Pringle manoeuvre
Liver resection conditions	Continuous Pringle manoeuvre (IP group) Intermittent Pringle manoeuvre (control group)	Selective hepatic vascular occlusion (IP group) Intermittent selective hepatic vascular occlusion (control group)
IP clamp time (min)	10	10
Reperfusion time (min)	10	10
Intermittent ischaemia/reperfusion time (min)	15/5	15/5
All resections ≥3 segments	Yes	No
Cirrhotic livers	No	No
Mortality		
IP	1 of 36	0 of 27
Control	0 of 37	0 of 27
Liver failure		
IP	0 of 36	–
Control	2 of 37	–
Postop. morbidity		
IP	14 of 36	9 of 27
Control	14 of 37	10 of 27
Hospital stay (days)*		
IP	15(10)	–
Control	13(9)	–
ITU stay (days)*		
IP	4(4)	–
Control	2(3)	–
Duration of operation (min)*		
IP	345(126)	211(31)
Control	300(116)	237(26)
Patients transfused		
IP	9 of 36	–
Control	11 of 37	–
Packed cells (units)*		
IP	1·7(1·8)	–
Control	2·3(2·0)	–
Blood loss (ml)*		
IP	426(350)	520(247)
Control	492(456)	720(220)
Prothrombin time (%)*		
IP	74(19)	49(6)†, 47(8)‡
Control	72(17)	48(5)†, 47(7)‡
Bilirubin (μmol/l)*		
IP	50(41)	56(6)†, 50(8)‡
Control	49(32)	52(6)†, 53(11)‡
AST (units/l)*		
IP	645(390)	735(216)†, 485(176)‡
Control	528(353)	680(115)†, 515(233)‡
ALT (units/l)*		
IP	414(360)	–
Control	522(377)	–

*Values are mean(s.d.). †Ischaemia time 40 min or more; ‡ischaemia time less than 40 min. IP, ischaemic preconditioning; ITU, intensive therapy unit; AST, aspartate aminotransferase; ALT, alanine aminotransferase.

CC. Of these, only three included an appropriate power calculation and none was powered on mortality. The outcome measures used were apoptotic rate in hepatocytes[24], all ischaemia-related complications including death[3], and probability of better outcome, which was determined from the following statement: 'The assumption was made that the difference between the two groups was such that the probability for a better outcome for one patient in the preconditioning group *versus* the patient in the control group was at least 0·7'[10]. There was no significant heterogeneity among the studies ($I^2 = 29$ per cent, $P = 0{\cdot}23$). No significant reduction in mortality following IP was

© 2013 British Journal of Surgery Society Ltd
Published by John Wiley & Sons Ltd

www.bjs.co.uk

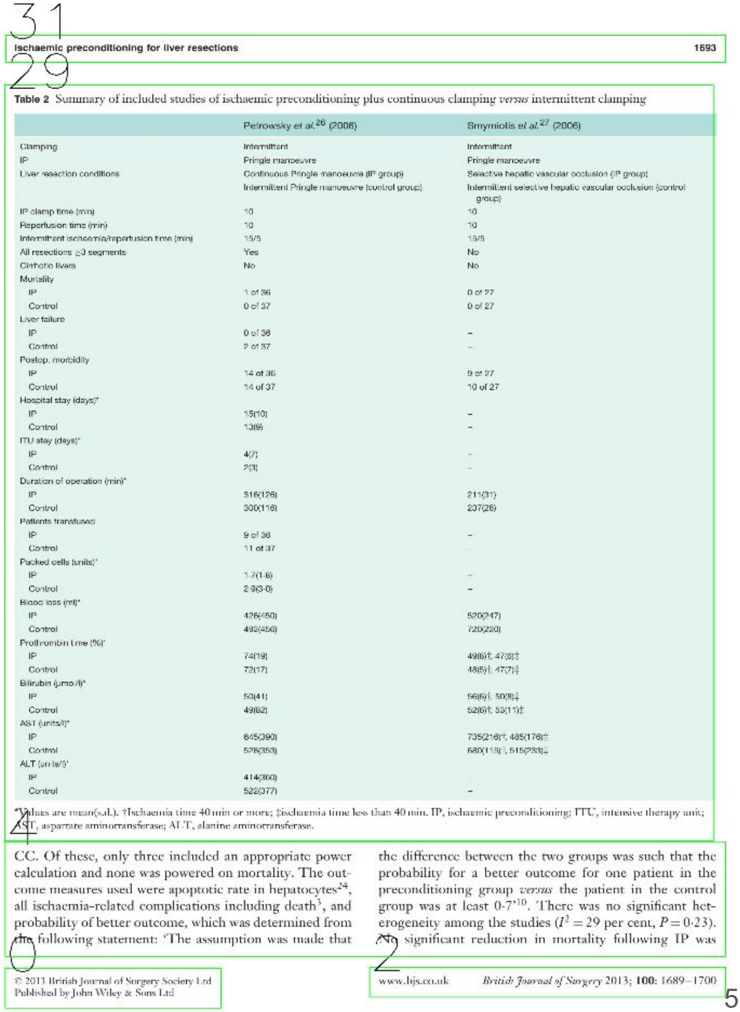
British Journal of Surgery 2013; **100:** 1689–1700

CV approaching - refining

Consolidate regions of interest

Filter out non-relevant content

Adjust kernel for more accurate segmentation



Ischaemic preconditioning for liver resections 1693

Table 2 Summary of included studies of ischaemic preconditioning plus continuous clamping *versus* intermittent clamping

	Petrowsky *et al.*[26] (2006)	Smymiolis *et al.*[27] (2006)
Clamping	Intermittent	Intermittent
IP	Pringle manoeuvre	Pringle manoeuvre
Liver resection conditions	Continuous Pringle manoeuvre (IP group) Intermittent Pringle manoeuvre (control group)	Selective hepatic vascular occlusion (IP group) Intermittent selective hepatic vascular occlusion (control group)
IP clamp time (min)	10	10
Reperfusion time (min)	10	10
Intermittent ischaemia/reperfusion time (min)	15/5	15/5
All resections ≥3 segments	Yes	No
Cirrhotic livers	No	No
Mortality		
IP	1 of 36	0 of 27
Control	0 of 37	0 of 27
Liver failure		
IP	0 of 36	–
Control	2 of 37	–
Postop. morbidity		
IP	14 of 36	9 of 27
Control	14 of 37	10 of 27
Hospital stay (days)*		
IP	15(10)	–
Control	13(9)	–
ITU stay (days)*		
IP	4(7)	–
Control	2(3)	–
Duration of operation (min)*		
IP	316(126)	211(31)
Control	300(116)	237(26)
Patients transfused		
IP	9 of 36	–
Control	11 of 37	–
Packed cells (units)*		
IP	1·7(1·6)	–
Control	2·9(3·0)	–
Blood loss (ml)*		
IP	426(450)	520(247)
Control	492(456)	720(220)
Prothrombin time (%)*		
IP	74(19)	49(6)†; 47(8)‡
Control	72(17)	48(5)†; 47(7)‡
Bilirubin (μmol/l)*		
IP	50(41)	56(6)†; 50(8)‡
Control	49(62)	52(6)†; 53(11)‡
AST (units/l)*		
IP	645(390)	735(216)†; 485(176)‡
Control	528(353)	680(115)†; 515(233)‡
ALT (units/l)*		
IP	414(360)	–
Control	522(377)	–

*Values are mean(s.d.). †Ischaemia time 40 min or more; ‡ischaemia time less than 40 min. IP, ischaemic preconditioning; ITU, intensive therapy unit; AST, aspartate aminotransferase; ALT, alanine aminotransferase.

CC. Of these, only three included an appropriate power calculation and none was powered on mortality. The outcome measures used were apoptotic rate in hepatocytes[24], all ischaemia-related complications including death[3], and probability of better outcome, which was determined from the following statement: 'The assumption was made that the difference between the two groups was such that the probability for a better outcome for one patient in the preconditioning group *versus* the patient in the control group was at least 0·7'[10]. There was no significant heterogeneity among the studies ($I^2 = 29$ per cent, $P = 0{\cdot}23$). No significant reduction in mortality following IP was

© 2013 British Journal of Surgery Society Ltd
Published by John Wiley & Sons Ltd

www.bjs.co.uk *British Journal of Surgery* 2013; **100**: 1689–1700

Further steps

Problems: very slow compared to traditional parsing (1-2 minutes per PDF compared to <10 seconds for traditional methods). Should only be used as last-resort backup.

Haven’t been deployed on HPC yet, quite a few packages needed.

There are definitely more development that we can do with this, e.g. kernel adjustment, separating text from tables, which can lead to interesting application (e.g. parsing tables through CV).

COVID-19 Indiviual Patient Dataset Update

Will and Zihan

Initial Tasks - Will

Goal: Label the sub journals of Nature and Lancet based on whether or not their articles have a Data Availability section.

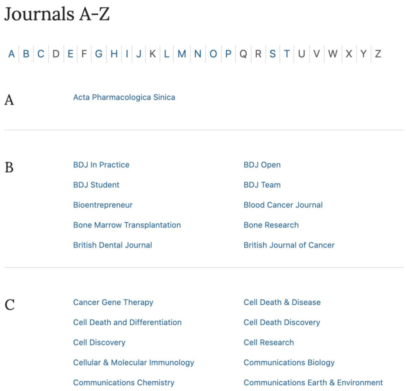
To do this I built a scraper for Nature and for Lancet, that would identify the sub journals and would search their COVID related articles for headers similar to Data Sharing/Data Availability.

Search

First I identified each Nature Journal as listed on https://www.nature.com/siteindex

In a Nature Search query, to filter by journal a unique journal identifier is required. This information was retrieved at this time as well.

After all ID's were found the script searched Nature (filtering by journal, article, date, covid related) and recorded the sub journals that had Data sections. A similar method was used for Lancet



Journals A-Z

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

A

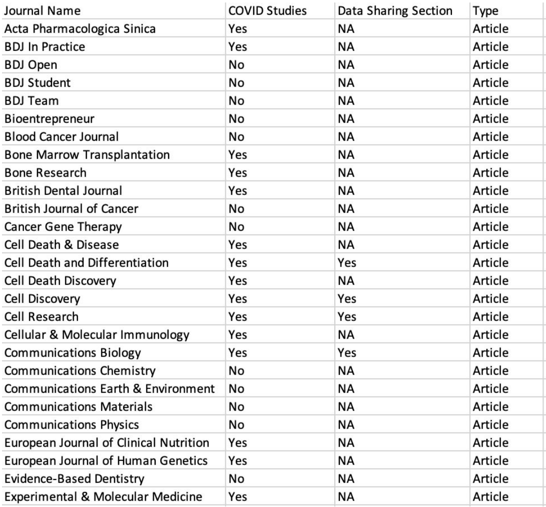
Acta Pharmacologica Sinica

B

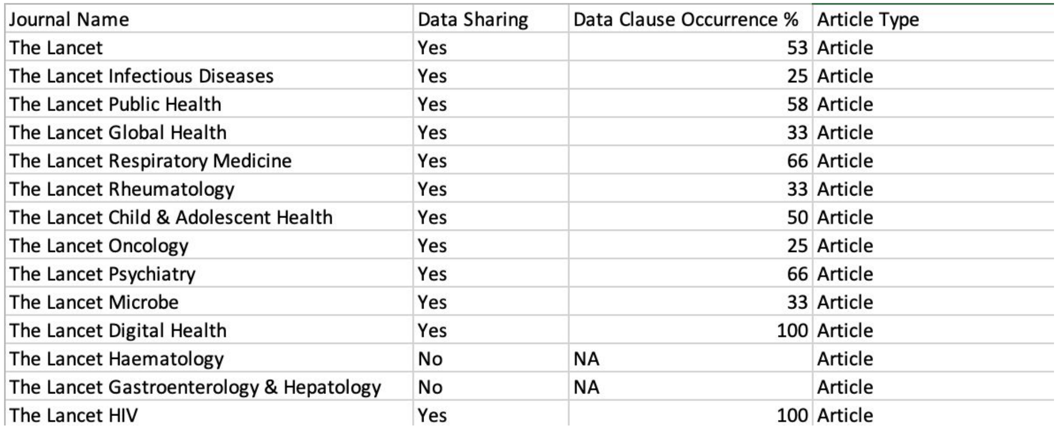
BDJ In Practice
BDJ Open
BDJ Student
BDJ Team
Bioentrepreneur
Blood Cancer Journal
Bone Marrow Transplantation
Bone Research
British Dental Journal
British Journal of Cancer

C

Cancer Gene Therapy
Cell Death & Disease
Cell Death and Differentiation
Cell Death Discovery
Cell Discovery
Cell Research
Cellular & Molecular Immunology
Communications Biology
Communications Chemistry
Communications Earth & Environment



Journal Name	COVID Studies	Data Sharing Section	Type
Acta Pharmacologica Sinica	Yes	NA	Article
BDJ In Practice	Yes	NA	Article
BDJ Open	No	NA	Article
BDJ Student	No	NA	Article
BDJ Team	No	NA	Article
Bioentrepreneur	No	NA	Article
Blood Cancer Journal	No	NA	Article
Bone Marrow Transplantation	Yes	NA	Article
Bone Research	Yes	NA	Article
British Dental Journal	Yes	NA	Article
British Journal of Cancer	No	NA	Article
Cancer Gene Therapy	No	NA	Article
Cell Death & Disease	Yes	NA	Article
Cell Death and Differentiation	Yes	Yes	Article
Cell Death Discovery	Yes	NA	Article
Cell Discovery	Yes	Yes	Article
Cell Research	Yes	Yes	Article
Cellular & Molecular Immunology	Yes	NA	Article
Communications Biology	Yes	Yes	Article
Communications Chemistry	No	NA	Article
Communications Earth & Environment	No	NA	Article
Communications Materials	No	NA	Article
Communications Physics	No	NA	Article
European Journal of Clinical Nutrition	Yes	NA	Article
European Journal of Human Genetics	Yes	NA	Article
Evidence-Based Dentistry	No	NA	Article
Experimental & Molecular Medicine	Yes	NA	Article



Journal Name	Data Sharing	Data Clause Occurrence %	Article Type
The Lancet	Yes	53	Article
The Lancet Infectious Diseases	Yes	25	Article
The Lancet Public Health	Yes	58	Article
The Lancet Global Health	Yes	33	Article
The Lancet Respiratory Medicine	Yes	66	Article
The Lancet Rheumatology	Yes	33	Article
The Lancet Child & Adolescent Health	Yes	50	Article
The Lancet Oncology	Yes	25	Article
The Lancet Psychiatry	Yes	66	Article
The Lancet Microbe	Yes	33	Article
The Lancet Digital Health	Yes	100	Article
The Lancet Haematology	No	NA	Article
The Lancet Gastroenterology & Hepatology	No	NA	Article
The Lancet HIV	Yes	100	Article

Downloads

42 sub journals were identified to have data availability sections, so I began downloading articles from these sub journals using COVID related terms and Biomarkers like D-Dimer,Procalcitonin,and CRP.

So far, most of these searches produce very few articles but most provide some sort of data. Many seem to be upon request, or links to outside sources that the researchers analyzed.

Collabovid

Secondary database includes COVID-19/SARS-CoV-2-related papers. Frequently updated.

No separate data file. Embedded in the text. Download the document for future reference.

Multiple pdf sources. Need to analyze independently.

Receive 72 outcomes related to these **3 biomarkers (CRP, D-Dimer and Procalcitonin)**.

Outline

1. Keyword search biomarkers (rather than semantic search)
    a. Focusing on Title and Abstract (Effective?)
2. Identify its source
    a. Elsevier (ScienceDirect): log-in requirement to obtain link. Referring to the codes in previous project.
    b. PudMed
        i. Find link “PudMed Central”, redirect to NCBI website
        ii. Find link “EMH Swiss Medical Publishers Ltd”, redirect to Swiss Medical Weekly website
        iii. Otherwise no full-text access
    c. medRxiv & bioRxiv & arXiv : PDF shows directly under the website. Store information in CSV file.
3. Download pdf using the downloadable link

Going Forward - Article Analysis

In order to determine similar articles the original presentation proposed utilizing Kmeans clustering with K=20.

Since the number of clusters, or groupings of articles, is not known Kmeans may not be our best choice. An algorithm like DBSCAN may be more preferable as noise does not affect the clustering and it will provide the best number of clusters rather than having to choose the number at random beforehand.

The similarity measure will ultimately depend on how we transform the articles but vectorizing them rather than using embeddings seems like the most reasonable approach.

Article Analysis

Once we identify the main clusters in the dataset we can use Latent Dirichlet Analysis to determine the topics used in each of the clusters as proposed in the original presentation.

With our current Journals, we may not have very many articles to work with and clustering may be lead to arbitrary results. We can either expand our downloads to include non Biomarker articles or search for more databases.