**Abstract**

The detriments of organizational silos have been studied extensively in business systems. Namely, organizational silos impede the flow of information between parts of an organization, resulting in wasted time and resources. In academic healthcare, the costs from organizational silos could potentially result in discoveries not being communicated to areas of healthcare where it could make a big difference. To begin the investigation of organizational silos in academic healthcare, we investigated siloing in the journals in the nephrology and transplant domains. To accelerate the investigation of silos, we leverage pretrained natural language processing models in our investigation, and provide a guideline for investigation of further silos.

**Methods**

* Data Collection
  + Software Packages: BioPython 1.84, Entrez Module, , selenium 4.21.0
  + A list of kidney and transplant journal with the top impact factors was compiled, (Figure 1) A search query was constructed to search PubMed for all the articles in each of the journals from the list (Figure 2).
  + Using the Entrez module from the BioPython package, there was an attempt made to extract the variables shown in Figure 3. Unavailable variables were left blank. Information for 622 articles ended up being extracted. Using the extracted publication types, the list of publication types in Figure 4 were filtered out, resulting in 492 articles being used.
  + For articles which did not have an abstract, the article text was extracted from the HTML of the journal site using Selenium 4.21. The HTML was accessed via institutional access to the journal website.
* Siloing
  + To determine department siloing, the unique named entities from each affiliation first needed to be extracted. This was done using the SpaCy package. Specifically, the default pretrained *en\_core\_web\_trf* model was used as it has the highest precision, recall, and F-Score of the current SpaCy models in the named-entity-recognition task. The model extracted the name-entities for each affiliation of each article. The departments were then filtered from the extracted named entities using regular expressions (RegEx). For each article, if the number of unique departments was equal to 1, the article was considered departmentally siloed.
  + To determine parent organization siloing, first a set of all unique affiliation strings across the entire set of articles was created. Next, the affiliations were separated by country. Then, a list of all universities in both the United States and the world was used to map the affiliations by matching the name of the university to the university name in the affiliation, if it was present. On a second pass, the unmapped affiliations were mapped as follows:
    - If the affiliation location is affiliated with a university, the university was considered the parent institution
    - If the affiliation location is a subsidiary of a company, the parent company is the parent organization
    - Otherwise, the affiliation location is its own parent institution

The second pass was done via manual research and RegEx.

Once the mapping of affiliations to parent institutions was created, themap was applied to each set of affiliation of each article, generating a set of unique parent organizations for each article. For each article, if the number of unique parent organizations was equal to 1, the article was considered parentally siloed.

* + To determine extreme siloing, for each article, if the article was parentally siloed and departmentally siloed, the article was considered extremely siloed.
* Article Classification
  + For each article, the articles were first classified into either being a kidney, transplant, or kidney and transplant (both) article. This was done using a keyword filter on the abstract text of the article, or the extracted article text if the abstract was not available.
  + Next, the topic of the article was determined using a zero-shot classification model. Zero-shot classification is a task in which a model, given a set of unique categories and the desired text to be classified, maps the text to the category which is most likely to correspond to the topic of the text. The model used is called *deberta-v3-base-tasksource-nli*, a state of the art zero-shot classifier based on the deberta-v3-base model from Microsoft. To reduce computational strain, the set of categories input to the model was dynamically switched depending on if the article was considered a kidney article, a transplant article or both, as seen in Figure 4.
* Analysis
  + For creation of the figures, csv files of the data, and analysis of the data, Pandas 1.5.3 was used.

**Citations**

* BioPython
* Entrez (Maybe)
* Pandas
* HugginingFace
  + Definitely deberta-tasksource
* Spacy
  + Probably also roberta
* SciPy

**Figures**

Figure 1. List of Kidney and Transplant Journals and Corresponding Impact Factors

Figure 2. Pubmed Query

Figure 3. Variables Extracted with Entrez

|  |
| --- |
| Title |
| PMID |
| Authors |
| Affiliations |
| Abstract |
| Journal Name |
| Grants and Funding |
| Conflict of Interest Statement |
| MeSH Terms |
| MeSH Term Descriptor IDs |
| Publication Type |
| Total Count of Authors |

























