in this issue >>>

Issue Three

A Quarterly Insight on the Services our Division Offers

DEB quarterly

Bulletpr007 Stats

An Introduction to Natural Language Processing

By: Duncan Vos, MS

Natural Language Processing (NLP) is an important and valuable tool that has been gaining use with the increase of available computing power. NLP involves the use of computers and automated algorithms to analyze and synthesize human language text. It enables us to uncover insights from human language text data.

Our interaction with NLP is more frequent than we may realize. NLP is seen with Google's suggested auto-completion of words or phrases when we begin a Google search. Another common instance can be seen when sending a text message, where grammatical errors are identified and reconciled. In healthcare research, NLP seems to be gaining use cases. It has been used with discharge summaries to detect adverse events (*Melton, JAMIA v12:4, p.448, 2005*) and was used to predict early psychiatric readmission (*Rumshisky, Nature TransPsych 6, e921, 2016*).

NLP makes use of algorithms, methodologies, and tools to analyze natural language text. It is a subfield of both computer science and linguistics. Generally, NLP is used to systematically break down language into smaller elemental pieces, which can then be used to assess relationships or associations. NLP can be used to study grammatical, syntactic, semantic, and sentiment structure of text. There are packages and open source code available to use NLP in R, Python, and SAS.

There are many capabilities of NLP. Topic modeling could be used to summarize themes in a collection of texts. Sentiment analysis could be used to identify the mood or emotional tones across a corpus of texts. Semantic analysis could be used to identify the most relevant topics in a text by identifying sets of synonyms (synsets) such as 'exercise' and 'workout', and then further quantifying the distance of other words used that are similar but less extreme (such as 'stretching', 'yoga') or more extreme (such as 'exertion') (*Gutterman, J Med Internet Res 2018;20(6)e231*).

These tasks involve tokenization, which is the process of identifying individual tokens (words, lemma, and/or punctuation) from a sequence of words. Furthermore, with these tasks, some form of a lexicon is typically used. A lexicon is essentially a dictionary of values stripped of the definitions and then each word contains quantitative or qualitative measures. For example, a lexicon that is used for sentiment analysis may have indication for positive words, negative words, anger words, fear words, and trust words.

The amount of natural human language text data is vast and constantly growing. From gardening forums to medical discharge notes to twitter feeds and voice-to-text internet searches. Up to 80% of medical record data is thought to be unstructured free text. Natural Language Processing is a quantitative analytical approach that could not only bolster a qualitative analysis by incorporating a quantitative component, but will play an important role in the future of healthcare.

Big Data Analytics for Healthcare Research

By: Ransome Eke, MD PhD
Physician Epidemiologist

Big data is an evolving term that describes any voluminous amount of structured, semi-structured and unstructured data that has the potential to be mined for information.

Big Data Analytics is the process of examining large and varied data sets to uncover hidden patterns and useful information that can help make more-informed decisions.

The 5 key properties of big data are - Volume, Velocity, Variety, Value and Veracity
Generally, the value and effectiveness of big data depends on:

- Human operators (Researcher) tasked with understanding the data
- Formulating proper queries to direct big data projects

Why is big data analytics important?

- To harness that data and extract value from it
- Cost effective
- Faster and better decision making
- Health care industries: on the horizon and important to take control of health information.
- Uncover hidden patterns, correlations, insights and intervention opportunities

BIG DATA ANALYTICS

Cost reduction

Faster, better decision making

New products & services

Example sources of "Big data"

- Hospital system: Electronic Health Record (EHR) e.g. Epic EHR
- Health Survey Databases: NHANES
- Health Administrative Databases:
    - Medicare Current Beneficiary Survey (MCBS)
    - Healthcare Cost and Utilization Project (HCUP)
- Institutions/Agencies:
    - Medical Information Mart for Intensive Care (MIMIC),
    - T1D Exchange
- Social media: YouTube, Facebook, twitter, Instagram, Reddit

How can big data be used in healthcare research?

- Predict health outcomes and create care plans
- Estimate of summary statistics: prevalence, incidence, mortality, etc.
- Stratify population by risk and develop insights into barriers faced
- Pilot data for grant proposals
- Power exploration
- Hypothesis generation & testing
- Secondary analysis ~ publications

Some limitations & Challenges of big data

- Non-Experimental
- Most are cross sectional
- May require special skills: statistical techniques & software usage
- Statistical issues to address
- May have a fee

How can the division of Epidemiology and Biostatistics help?
The division has the resources and expertise to help both clinicians and non-clinicians work through the process of using big data techniques for research. These processes include – identifying and formulating research questions, identifying data requirements, pre-processing of data, performing analytics of the data and visualizing the data.

Furthermore, the division has procured some big data to boost research at WMed. We currently have data from Healthcare Cost and Utilization Project (HCUP) for various population, settings and years. These include the National (Nationwide) Inpatient Sample (NIS), Kids' Inpatient Database (KID), Nationwide Readmissions Database (NRD) and Nationwide Emergency Department Sample (NEDS).

If you need more information on how to navigate the big data analytic process, feel free to visit our division or contact me at ransome.eke@med.wmich.edu

EPI
CORNER

Writers Block
By: Dr. Laura Bauler, PhD

Navigating the Publication Process-Part 1: Finding the Best Journal

By: Laura Bauler, PhD
Medical Editor

There comes a time in every research project when you have to start thinking about where you are going to publish your findings. Whether you are submitting the article for the first time or resubmitting to a new journal, identifying the best location for your manuscript can be a challenge. There are a number of factors you should consider before selecting a journal including: indexing, journal ranking, publication frequency, predatory journal status, cost of publication, article acceptance rate, and the journals' author guidelines.

1. **Indexing**: Journal databases contain articles from journals that they index. Databases such as PubMed, SCOPUS, or EMBASE, all have certain journals that they index, meaning articles can be searched for and identified by users of those databases. If you plan to publish in a journal that is NOT indexed, how will readers find your article? Your article can only be cited if it can be found and read by others.
2. **Journal Ranking**: A number of journal ranking tools are available that rank journals based largely upon the number of citations a certain journals' articles receive. While a highly debated topic as to which ranking algorithm is best, it is largely agreed that citations are a mark of the quality of an article.
3. **Publication Frequency**: How frequently does the journal you are interested in publishing in get published, biweekly, monthly, bi monthly, quarterly, yearly.
4. **Predatory Journal status**: Predatory journals are journals that do NOT adhere to the scientific standards of peer review utilized by scientists. They publish, typically for a fee, articles that either do not get subjected to peer review or undergo a weak peer review process. Identifying these predatory journals can be a challenge. http://thinkchecksubmit.org/check/
5. **Cost of publication:** Journals vary widely in cost. Some open access journals rely upon authors to fund that publication and may cost several thousand dollars, while many traditional subscription based journals have no author fees at all for publication. Other journals may be a mix of the two with journal publication fees used to offset the costs of the publication, such as page charges, color print charges or article processing charges.
6. **Article acceptance rate:** The percent of articles that are submitted to a journal that get accepted. Submitting to journals with a higher article acceptance rate will improve your chances of making it through the publication process. However journals with a higher ranking will likely have a lower acceptance rate, due to the higher number of submissions despite a steady number of publications.
7. **Author guidelines:** Every journal develops author submission guidelines that provide information about the types of articles the journal accepts and how to format your manuscript for that journal. Make sure they are interested in the type of manuscript you plan to submit.

In the table below you will find several methods that can be used to identify the best journal for your manuscript.

Journal Selection Method	**Tips**
Current manuscript citations	Within the reference list for the manuscript you are currently writing there are probably several references that came from the same journal. Due to the similarity in topic and nature of your manuscript, there is a good chance that these journals and their readership would be a good audience for your manuscript.
Biosemantics tools	1. JANE (Journal, Author, Name, Estimator). Jane is a tool that identifies and matches keywords from your abstract to literature already published and indexed in PUBMED. http://jane.biosemantics.org/. 2. Elsevier Journal Matcher. This tool matches your abstract to potential journals published by Elsevier. https://journalfinder.elsevier.com/. 3. Springer Journal suggester. This tools matches your abstract to relevant journals published by Springer and BMC. https://journalsuggester.springer.com/. 4. Edanz Journal Selector. Developed by Edanz editing company, searches over 28000 journals. https://www.edanzediting.com/journal-selector 5. Journal Guide by the American Journal Experts. Utilizes a proprietary algorithm to search more than 46000 journals. https://www.journalguide.com/
Colleagues	Discuss options with colleagues and mentors who know the field, where a particular article should be submitted.
Journal ranking lists	1. Journal Citation Reports: Journal impact factor is a ratio of citations and recently citable items published. http://libguides.med.wmich.edu/az.php?a=j 2. SCIMAGO Journal Rank (SJR): Weighs the prestige of a journal based upon where its articles are cited https://www.scimagojr.com/journalrank.php 3. SCOPUS cite score: # citations a journal receives in 1 year for articles published in the last 3 years. Divided by the number of documents indexed in SCOPUS during those 3 years. https://www.scopus.com/sources.uri?zone=TopNavBar&origin=sbrowse

More information can be found at the WMed Academic Scholarship and Writing Guide.
http://libguides.med.wmich.edu/scholarshipwriting/scholarlyPublishing/sources

Need help? Contact: Laura Bauler PhD, or the WMed library staff.

Data Bytes

Anonymous and Confidential
What’s the Difference and Why Does it Matter?

By: Dan Foley
Database Specialist

Problem: A distressing number of first year medical students still think you should put something in a seizing persons mouth to prevent them from swallowing their tongue.

Solution: Examine the impact a one hour first aid course has on your first year student’s knowledge of how to treat seizures. Create a quiz, with questions pertaining to what actions to take when you find someone unresponsive and convulsing. You want the most honest answers, so you let your subjects know that their responses are *anonymous* and do not ask for the student’s name. You give this test before and after the class.

You find the average score before the class, and the average score after the class. The assumptions for a Student’s t-test appear to hold true. You perform your Student’s t-test and you find that you have a significant result! High-fives all around and you prepare your poster for research day.

The problem, is that you are doing it wrong.

Your study population data (before and after scores) are correlated. They are the same people, taking the same test. What you need is a paired t-test, and for that you need to link individual before and after scores. You need an identifier.

You can create identifiers and still have an anonymous data collection scheme. In the above example, if you pre-numbered both the before and after tests, and handed them out together, you could link the before and after scores using the test number. No way to link the test number to the participant, so the data remains anonymous.

In many cases, when conducting research, similar anonymous labelling schemes just aren’t possible. To properly conduct your research, you will have to use an identifying variable such as e-mail address, name, or MRN. In these cases, what you need is *confidential* data collection. Confidential data collection requires careful forethought, execution, and oversight. At the Division of Epidemiology and Biostatistics, our data management team is excited to work with you to ensure these requirements are met.

WMed provides support for a REDCap research database, which has extensive capabilities to protect your data and your study participant’s confidentiality. Through a variety of strategies including self-generated identifiers, date shifted data, de-identified databases, LDAP authentication, and user logging; we can provide anonymous or confidential data collection solutions for your research needs.

Q: *How do we contact you for project assistance?*

A: *That’s easy, visit the [Divison Webpage] and submit a request or just reach out to us at epibio@med.wmich.edu*

final thoughts...

2019 KALAMAZOO COMMUNITY RESEARCH DAY

Tuesday, April 16th (Posters) & Wednesday, April 17th (Oral)

Poster Presentation Schedule: TBD

Oral Presentation Schedule: 8:00 a.m. – 12:45 p.m. (Registration starts at 8 a.m.)

WMed W.E. Upjohn Campus at 300 Portage Street, Kalamazoo, MI 49007

Pre-register for the event here >>> ***Research Day Pre-Registration Form***

Based on the feedback we have received, we have made some drastic changes based on your feedback as we begin to prepare for the 2019 Research Day event:

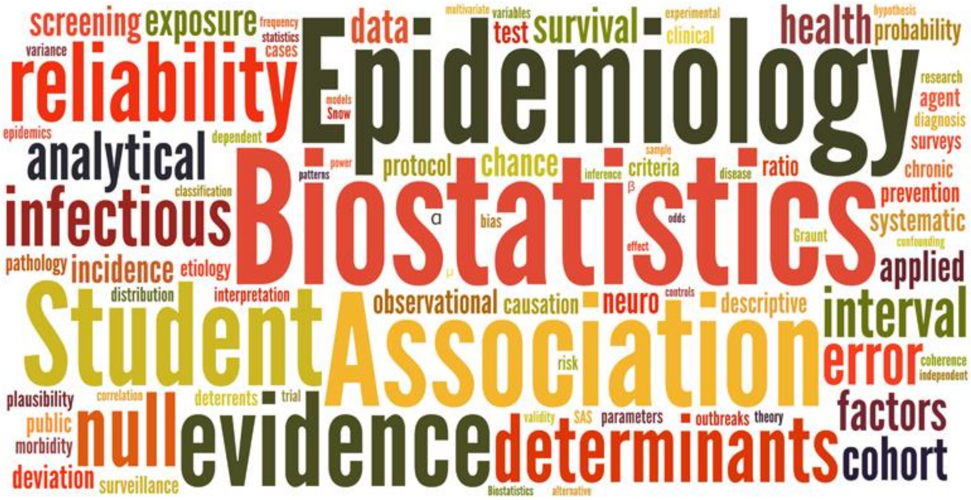
- The venue will now be at WMed W.E. Upjohn Campus.
- Poster presentations and oral presentations will now occur on separate days. Refreshments will be provided on both days.
- Posters will now be hung up at the venue for the entire week of April 15th, 2019.
- The abstract submission form and the post-event evaluation are being updated so they are more user friendly.
- A new poster printing process will be put into place, please watch out for updates regarding that process.

DEB

Epi & Bio
DIVISION

1000 Oakland Drive
Kalamazoo, MI 49008

Editor: Leah Bader
Contact us at 269-337-4609
Epi & Bio Website
epibio@med.wmich.edu



coming soon >>>

Important Deadlines for Research Day

Wednesday, January 2 – Online form opens for abstract submission at 8:00 a.m.

Wednesday, February 6 – Deadline for abstract submission at 11:30 p.m. (EST)

Friday, February 8 – Scoring begins

Friday, February 22 – Complete judging of abstracts

Friday, March 1 – Authors notified of results

Monday, April 1 – Poster submission deadline – MUST submit posters to researchday@med.wmich.edu

Monday, April 15 – Oral PowerPoint presentation deadline – MUST submit by 12:00 p.m. (EST)

Tuesday, April 16, 2019 – 2019 Research Day Poster Presentations

Wednesday, April 17, 2019 – 2019 Research Day Oral Presentations & Keynote Speaker (Keynote Speaker is TBD)

Research Day related questions can be directed to Leah Bader at researchday@med.wmich.edu