Using Natural Language Processing (NLP) to Streamline Literature Selection for Meta-Analysis (MA)

Jenny Ding¹, Youfang Cao², Sean Hayes¹, Gregory Bryman², Kelly Yee¹

¹Quantitative Pharmacology & Pharmacometrics, Merck & Co. Inc.
²Pharmacometrics, Eisai Co., Ltd.
³Research & Development Sciences IT - Data Science & Scientific Informatics, Merck & Co. Inc.

IQ Consortium AI/ML Workshop
September 15, 2022
Meta-Analysis PRISMA Flowchart

- Meta-Analysis leverages published evidences to inform discovery and clinical decision making
- However, screening and selecting relevant literature from PubMed and other databases are resource/time-consuming

1056 potentially relevant publications identified
  - PubMed Medline (n = 891)
  - COCHRANE (n = 165)

217 publications retrieved & screened in full text for inclusion in DB

839 excluded on initial abstract screening
  - Non-randomized study, no treatment details etc.

159 publications included in initial DB
  - PubMed Medline (n = 149)
  - COCHRANE (n = 10)

58 excluded after full text screening
  - Duplicates, no endpoints etc

87 publications included in final analysis DB
  - Providing results from 89 clinical trials

72 excluded to reduce to analysis data base
  - Drugs/Endpoints not in final analysis drug list

Plock et. al, 2017 Clin Pharm

PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses)
1056 potentially relevant publications identified from keyword search

217 publications retrieved & screened in full text

839 excluded on initial abstract screening
- Non-randomized study, no treatment details etc.

58 excluded after full text screening
- Duplicates, no endpoints, etc.

159 publications included in initial DB

159 publications included in initial DB

87 publications included in final analysis DB

72 excluded due to drug/endpoints not in final analysis list

72 excluded due to drug/endpoints not in final analysis list

87 publications included in final analysis DB

87 publications included in final analysis DB

Cleaned and formatted data ready for analysis

Cleaned and formatted data ready for analysis

NLP Advantages
- A few minutes of run time vs months of manual curation
- $5 computing cost vs 6 R3 FTE-months
- Streamlined process, less bias

Human

Human

Named Entity Recognition

Named Entity Recognition

Entity extraction
- PICO (patient, intervention, comparison, outcome)
- PK Parameters

Entity extraction
- PICO (patient, intervention, comparison, outcome)
- PK Parameters

Proprietary

Natural Language Processing for Automated Literature Selection
Exploration of NLP methods show Unsatisfactory Performance

Raw PubMed search: 32% of abstracts selected with the cut are relevant

Cosine Similarity with Query

 Ranked w/ Facebook **BioSentVec**: 38% relevant abstracts

Ranked w/ Google **Universal Sentence Encoder**: 37% relevant abstracts

Ranked w/ **TF/IDF** (term frequency-inverse document frequency): 55% relevant abstracts

Total N=412 abstracts

Use case is from a "NeuroPain" MBMA.

Improvements with domain-specific encoding
Transformer Models are Revolutionizing Biomedicine

**2017:** Transformers

---

**2018:** BERT

---

**2020:** PubMedBERT

---

**2021:** AlphaFold (BERT-based model) revolutionized protein 3D structure prediction

---

*It will change everything*: DeepMind’s AI makes gigantic leap in solving protein structures

Google's deep-learning program for determining the 3D shapes of proteins stands to transform biology, say scientists.
Transformer-based NLP Framework for MBMA Abstract Ranking

Relevant (R)
In Merck's meta-analysis database

All search results from PubMed using previous meta-analysis queries

Irrelevant (I)

Model pretrained by Microsoft


Transformers can understand Language Context

- Transformers (like BERT) have **attention** mechanisms that can learn **semantics** instead of only word frequency (TF-IDF), which is insufficient to capture long-term dependencies in sequences.

<table>
<thead>
<tr>
<th>Layer: 0</th>
<th>Attention: All</th>
</tr>
</thead>
</table>

Example of attention as shown from BertViz

T-SNE visualization of tokens selected (R) and not selected (I) shows high overlap.

World cloud of abstracts selected (R) and not selected (I) for MBMA is hard to differentiate.
Generalization: PubMedBERT can Predict Diseases not in Training Dataset

- Task 1: Leave-1-disease-out cross validation
- Train a model on 13 diseases and test model on the left-out disease
  - E.g., train on HCV to Endometriosis, test on Asthma

<table>
<thead>
<tr>
<th>Disease</th>
<th>Total PubMed results</th>
<th>Human selected papers</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asthma</td>
<td>60</td>
<td>33</td>
<td>100%</td>
<td>82%</td>
</tr>
<tr>
<td>HCV</td>
<td>1343</td>
<td>164</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Psoriasis</td>
<td>856</td>
<td>125</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RA</td>
<td>2062</td>
<td>208</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neuro pain</td>
<td>412</td>
<td>99</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T2 diabetes</td>
<td>8994</td>
<td>921</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T1 diabetes</td>
<td>2319</td>
<td>148</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Osteoporosis</td>
<td>2475</td>
<td>171</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Schizophrenia</td>
<td>3161</td>
<td>239</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NASH</td>
<td>1647</td>
<td>117</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lung cancer</td>
<td>1577</td>
<td>374</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dyslipidemia</td>
<td>3189</td>
<td>384</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grass pollen allergy</td>
<td>398</td>
<td>59</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Endometriosis</td>
<td>486</td>
<td>117</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total/Mean</td>
<td>28979</td>
<td>3159</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Recall: %Captured by Model out of All True Positives

Precision: %True Positives out of All Predicted to be Positives
### Generalization: PubMedBERT can Predict Diseases not in Training Dataset

<table>
<thead>
<tr>
<th>Disease</th>
<th>Total PubMed results</th>
<th>Human selected papers</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asthma</td>
<td>60</td>
<td>33</td>
<td>100%</td>
<td>82%</td>
</tr>
<tr>
<td>HCV</td>
<td>1343</td>
<td>164</td>
<td>93%</td>
<td>28%</td>
</tr>
<tr>
<td>Psoriasis</td>
<td>856</td>
<td>125</td>
<td>84%</td>
<td>31%</td>
</tr>
<tr>
<td>RA</td>
<td>2062</td>
<td>208</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neuro pain</td>
<td>412</td>
<td>99</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T2 diabetes</td>
<td>8994</td>
<td>921</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T1 diabetes</td>
<td>2319</td>
<td>148</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Osteoporosis</td>
<td>2475</td>
<td>171</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Schizophrenia</td>
<td>3161</td>
<td>239</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NASH</td>
<td>1647</td>
<td>117</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lung cancer</td>
<td>1577</td>
<td>374</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dyslipidemia</td>
<td>3189</td>
<td>384</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grass pollen allergy</td>
<td>398</td>
<td>59</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Endometriosis</td>
<td>486</td>
<td>117</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total/Mean</strong></td>
<td><strong>28979</strong></td>
<td><strong>3159</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Task 1:** Leave-1-disease-out cross validation
- Train a model on 13 diseases and test model on the left-out disease
  - E.g., train on *Asthma* to *Endometriosis* (EXCLUDING *HCV*), then test on *HCV*
- Repeat (retrain 12 other models), so each disease has a chance to be the test set
Generalization: PubMedBERT can Predict Diseases not in Training Dataset

<table>
<thead>
<tr>
<th>Disease</th>
<th>Total PubMed results</th>
<th>Human selected papers</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asthma</td>
<td>60</td>
<td>33</td>
<td>100%</td>
<td>82%</td>
</tr>
<tr>
<td>HCV</td>
<td>1343</td>
<td>164</td>
<td>93%</td>
<td>28%</td>
</tr>
<tr>
<td>Psoriasis</td>
<td>856</td>
<td>125</td>
<td>84%</td>
<td>31%</td>
</tr>
<tr>
<td>RA</td>
<td>2062</td>
<td>208</td>
<td>93%</td>
<td>24%</td>
</tr>
<tr>
<td>Neuro pain</td>
<td>412</td>
<td>99</td>
<td>78%</td>
<td>47%</td>
</tr>
<tr>
<td>T2 diabetes</td>
<td>8994</td>
<td>921</td>
<td>94%</td>
<td>24%</td>
</tr>
<tr>
<td>T1 diabetes</td>
<td>2319</td>
<td>148</td>
<td>96%</td>
<td>15%</td>
</tr>
<tr>
<td>Osteoporosis</td>
<td>2475</td>
<td>171</td>
<td>88%</td>
<td>15%</td>
</tr>
<tr>
<td>Schizophrenia</td>
<td>3161</td>
<td>239</td>
<td>92%</td>
<td>17%</td>
</tr>
<tr>
<td>NASH</td>
<td>1647</td>
<td>117</td>
<td>83%</td>
<td>15%</td>
</tr>
<tr>
<td>Lung cancer</td>
<td>1577</td>
<td>374</td>
<td>75%</td>
<td>45%</td>
</tr>
<tr>
<td>Dyslipidemia</td>
<td>3189</td>
<td>384</td>
<td>75%</td>
<td>23%</td>
</tr>
<tr>
<td>Grass pollen allergy</td>
<td>398</td>
<td>59</td>
<td>66%</td>
<td>25%</td>
</tr>
<tr>
<td>Endometriosis</td>
<td>486</td>
<td>117</td>
<td>74%</td>
<td>44%</td>
</tr>
<tr>
<td><strong>Total/Mean</strong></td>
<td><strong>28979</strong></td>
<td><strong>3159</strong></td>
<td><strong>85%</strong></td>
<td><strong>31%</strong></td>
</tr>
</tbody>
</table>

All leave-one-disease-out cross validation results

Smaller dataset show higher variability in outcomes
Proprietary

Generalization: Model trained on Historical Data can classify New Publications

- Task 2: Train a single 14-disease model on previous 3-year data of each disease and test on most recent 3-year
  - E.g., train on 2002-2006 data for asthma and on 2003-2010 data for HCV and ...(12 other diseases)
  - Then, test on 2007-2010 asthma and on 2011-2014 HCV abstracts and ...(12 other diseases)

Index of abstract (ordered by predicted score)

<table>
<thead>
<tr>
<th>Result</th>
<th>Pred Pos</th>
<th>Pred Neg</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Pos (R)</td>
<td>5570</td>
<td>1096</td>
</tr>
<tr>
<td>True Neg (I)</td>
<td>127</td>
<td>422</td>
</tr>
</tbody>
</table>

Recall = 77%

Precision = 28%
Pilot on Endemic SARS-CoV-2 Show Promising Results

Initial pilot on endemic SARS-CoV-2 vaccine
- Both clinical and non-clinical
- Outcomes include viral replication/titer, antibodies, hospitalization, etc.
- Purpose is to build a model that can use animal data to predict vaccine success

Time

779 literature from keyword search

346 literature selected by rule-based filter

16 literature selected by Merck experts (Rate limiting step)

<table>
<thead>
<tr>
<th>Result</th>
<th>Pred Pos</th>
<th>Pred Neg</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Pos (R)</td>
<td>14</td>
<td>2</td>
</tr>
<tr>
<td>True Neg (I)</td>
<td>60</td>
<td>703</td>
</tr>
</tbody>
</table>

Recall = 87.5%
Precision = 19%

- Used a non-stringent cutoff of 0.01 to catch more potentially relevant papers
- 74/799 abstracts selected, 90.5% reduction versus 44.4% by quick scan
- Need to reduce false negative ratio (12.5%)
Next Step: Entity Extraction for more interpretable features and enhanced flexibility

- More inclusive and streamlined alternative to manual curation
- Example of an abstract ranked highly by algorithm but missed/left out by manual selection

To compare the efficacy and safety of liraglutide versus sitagliptin as add-on to metformin after 26 weeks of treatment in Chinese patients with type 2 diabetes mellitus (T2DM). This 26-week open-label, active comparator trial (NCT02008682) randomized patients (aged 18-80 years) with T2DM inadequately controlled with metformin [glycated haemoglobin (HbA1c) 7.0-10.0% (53-86 mmol/mol)] 1 : 1 to once-daily subcutaneously administered liraglutide 1.8 mg (n = 184)...The primary endpoint was change in HbA1c from baseline to week 26. Liraglutide was superior to sitagliptin in reducing HbA1c from baseline [8.1% (65 mmol/mol)] to 26 weeks, as evidenced by estimated mean HbA1c change of -1.65% (-18.07 mmol/mol) versus -0.98% (-10.72 mmol/mol)...More patients receiving liraglutide (76.5%) than sitagliptin (52.6%) achieved the HbA1c target...
Conclusions

• **Summary:**
  - BERT-based NLP methods outperform traditional NLP methods (e.g., TF-IDF)
  - Potentially a cheaper and quicker alternative

• **Leveraged state-of-the-art biomedical-specific NLP model:**
  - Fine-tuned a neural network classifier on top of PubMedBERT model using internal MBMA data

• **In test set used to date, generalized to unseen disease and unseen (non-training set) abstracts**
  - 85% Recall in capturing top-ranked abstracts of unseen diseases
  - 77% Recall in predicting abstracts published downstream of training data

• **Future efforts:**
  - Conduct more pilot testing over different therapeutic areas
  - Reduce false negative ratio by further fine-tuning
  - Expand functionality based on literature curation needs; e.g., entity recognition