View Reviews

Paper ID
100

Paper Title
Mapping Unstructured Medication Orders to Standardised AEOLUS Names: Methods and Applications

Track Name
Research Papers

Reviewer #3

Questions

1. Summarize the paper and its main contributions
This paper builds a pipeline to map unstructured medication orders to standardized AEOLUS names, and demonstrates that standardized medication names can improve the generalizability of predictive models.

2. What generalizable insights did the authors claim they are making to machine learning in the context of healthcare?
As the authors claimed, they show that "mapping drug brand names to generic concept names can improve the accuracy of a standard ML benchmark task". Their work "presents the first open-access standardized set of generic drug concepts that can be easily used by other ML researchers to compare between independent datasets or to build generalisable models". Lastly, "the mapping enables the inclusion of additional, relevant information like active-ingredients, drug roles/categories, by easily integrating with existing drug knowledge databases that contain such information for generic drugs".

3. Were the claims of these insights supported in the body of the paper?
The claims were supported in the paper.

4. Please provide detailed comments, including strengths and weaknesses of the paper.
This paper tackles an important problem of using unstructured medication data in ML research. However, the mapping method used in this paper is not new, and thus lacking technical novelty. This paper is well written with convincing results, but unfortunately does not suit MLHC.

5. Is your main expertise on the clinical or computational side (or both)?
Computational

6. Please enter your overall assessment for this paper.
weak reject

Reviewer #4

Questions

1. Summarize the paper and its main contributions
The authors propose an approach for mapping medication names to standard generic names taken from the AEOLUS database. They demonstrate the application of the proposed mapping pipeline to mortality prediction in MIMIC-III dataset as well as a private dataset. They show that the mapping to AEOLUS increases the comparability of the two distinct datasets and can lead to better generalizability of ML models to new datasets.

2. What generalizable insights did the authors claim they are making to machine learning in the context of healthcare?
The paper suggests mapping drug brand names to generic concept names to improve the accuracy of standard ML benchmark tasks. The authors also present an open-access standardized set of generic drug

concepts to further facilitate their use in other ML tasks. The authors claim that the mapping allows the integration of additional information like active-ingredients, drug roles/categories using existing drug knowledge databases.

3. Were the claims of these insights supported in the body of the paper?
According to Table 5, their proposed mapping provides only a limited advantage when the training and the testing data are taken from the same dataset. In fact, the precision is decreased by ~1% when AEOLUS names are used in this scenario. The authors do show significant improvement in the performance when the training and test data are taken from different datasets. Yet, the performance is still significantly lower compared to the scenario where the training data is taken from the same dataset. Specifically, the use of Training Set B, which is significantly smaller than Training Set A, provides significantly better performance on Testing set B. Therefore, using Training Set A (with AEOLUS names) seems not practical. I think that a more practical experimental setting would be using both datasets for training. For example, using Training Sets A+B and Testing set B. This way the advantages of AEOLUS names can be evaluated in a transfer learning setting.

Also, I feel that the methodological contribution in the context of this conference is limited. Specifically, the generation of AEOLUS names is mainly based on standard data preprocessing stages and further requires manual work by clinicians. Moreover, clinical domain expertise, was further required for >10% for distinct generic drugs in both datasets A and B after applying the proposed mapping.

4. Please provide detailed comments, including strengths and weaknesses of the paper.
The strength of the paper is in presenting an approach to map medication names to standard generic names taken from the AEOLUS database. The authors show that this approach leads to an improved performance for the application of mortality prediction in the scenario where training and test data are taken from two different datasets. This mapping can be potentially used for other datasets.

However, both the methodological contribution and the experimental results are limited as noted above

5. Is your main expertise on the clinical or computational side (or both)?
Computational

6. Please enter your overall assessment for this paper.
weak reject

Reviewer #5

Questions

1. Summarize the paper and its main contributions
This paper maps free-text drug administrations to a previously published ontology of drug names.

2. What generalizable insights did the authors claim they are making to machine learning in the context of healthcare?
They assert their algorithm can be used to map drug names to a standard set of concepts across distinct databases.

3. Were the claims of these insights supported in the body of the paper?
Not entirely. They demonstrate that the mapping does work as it improves the performance of a mortality prediction model. However, this is an insufficient validation as (1) we do not know how well the model should perform, and (2) it is not an interpretable assessment of the quality of the mapping, it merely shows that it works in some cases.

4. Please provide detailed comments, including strengths and weaknesses of the paper.
Matching strings using edit distance is a very old and well established technique. In the specific context of concept mapping, OHDSI have a tool which matches using edit distance - WhiteRabbit. I mention this as you appear to use OHDSI tools already. It would be useful to discuss the added benefit of your approach.

Naive string matching on medical terms is risky as there are very distinct conditions with very low edit distance

(e.g. "AFIB" and "VFIB"). The use of edit distance also does not incorporate varying levels of specificity, e.g. Imatinib Mesylate is identical to Imatinib but has large edit distance. The edit distance additionally fails to handle synonyms; e.g. Imatinib is identical to Gleevec. These are well known issues in the literature that are not considered in this paper.

The primary issue with this paper is the assessment of the approach. The baseline of using distinct medication order phrases as a single feature is unrealistic. A better baseline would be the use of a bag-of-words model, where the component words are individual features. That is, "humalog insulin" would become two features, "humalog" and "insulin".

Nevertheless, validation with a downstream task is a common approach in clinical machine learning. It is useful here, but incomplete. Given the importance of accurately capturing a patient's medications in a clinical study, there must be some assessment of the accuracy of the mapping method itself. I would propose two approaches. First, clinical experts could manually label the records. There are a total of ~10,000 unique drug names from prescriptions/inputevents_mv in MIMIC-III and validating 10,000 string pairs (or a subset) is not an unreasonably burdensome task for a few individuals. Alternatively, 80% of the medications in MIMIC-III are associated with a National Drug Code (NDC), which uniquely represents the drug and can be used to directly map to RxNorm and the concept. The known mapping could be used as a labeler to assess the text-only based approach.

I appreciate the open source approach. Is there any effort to make the UAE dataset available? It seems as though the medications could be posted on to GitHub with no risk to patient privacy, and it would improve reproducibility of the paper.

Other comments:

- I highly doubt you are the first to propose standardized generic drug concepts. Strong jusification is needed for this claim. Better to remove it.
- Examples of mapped medications with their original, processed, and mapped forms from both datasets would be useful for qualitative interpretation of the approach.
- You state that MIMIC-III merges two ICU systems, but the database also contains hospital wide data. In particular, the prescriptions table is sourced from a custom EHR.
- INPUTEVENTS_MV contains intravenous administrations, which is a component of but not identical to patient fluid intake.
- The mapping approach groups drugs regardless of route or dose - this is a limitation as researchers will almost always require this information to appropriately identify the relevant medications (e.g. gentamicin as a cream is very different indication to IV gentamicin).

5. Is your main expertise on the clinical or computational side (or both)?
Computational

6. Please enter your overall assessment for this paper.
strong reject

Reviewer #6

Questions

1. Summarize the paper and its main contributions
This paper proposed a software pipeline to map the raw medication orders in different EHR systems to generic drug names in AEOLUS database. The authors showed that after this mapping, the intersection of drugs between different EHR systems significantly increased, and the knowledge learnt from the mortality prediction task on one EHR system could be transferred to another EHR system with better performance.

2. What generalizable insights did the authors claim they are making to machine learning in the context of healthcare?
The authors claimed that the clinical data cleaning pipeline which matches the raw drug names in various local EHR systems to a universally standard naming system could help make the cleaned clinical data more consistent and improve the generalizability of machine learning models across different medical centers.

3. Were the claims of these insights supported in the body of the paper?
The authors showed that
1) the commonalities between the medical orders in two EHR systems increased significantly after the drug name matching, thus the cleaned datasets are more consistent with each other compared to pre-cleaning. However, due to the concerns stated in Weaknesses 2), I think this increased intersection is not strong enough in supporting the authors' claim.
2) the AUC of the transfer learning of mortality prediction increased after the cleaning and matching of the raw drug names.

4. Please provide detailed comments, including strengths and weaknesses of the paper.
Strengths:

1) The authors released their code, thus this pipeline can be reproduced and used by other researchers.
2) The writing is quite clear overall in spite of some minor typos and grammatical issues.
3) The authors informed researchers in this community of the AEOLUS database, which might be useful in the future research in similar domains.

Weaknesses:

1) The technical significance of this paper is not sufficient. Basically the authors described some quite commonly used text cleaning and word matching steps. Their pipeline does not deal with multiple drug names in one medical order, or takes advantage of the context / dependencies between words. I believe there exist quite a few techniques in the NLP domain which can make this pipeline more sophisticated.
2) The evaluation of the pipeline is not sufficient. I am a little concerned whether the DL distance is a good metric to measure the similarity between two drugs, e.g. could two drugs have a same ingredient (a long word) and a different other ingredient (a very short word), thus they share a same prefix and satisfy the threshold t but are totally different? If this case exists, there might be quite a few mis-matchings. If there can be some analysis of clinicians' reviewing of the matching results (even on a randomly chosen small subset considering the time and efforts), the quality of this matching pipeline will be more convincing.
3) What are the features used in the logistic regression model for mortality prediction? Is it a bag-of-words like feature set?
4) In spite of the improvement in the AUC for transfer learning, I am concerned that the AUCs and precisions are still not good enough for practical use, and the positive/negative class distributions in both datasets need to be provided to assess the results. Also, no cross-validation or bootstrapping like methods were implemented thus the statistical significance of this improvement in the AUC cannot be demonstrated.

5. Is your main expertise on the clinical or computational side (or both)?
Computational

6. Please enter your overall assessment for this paper.
strong reject