Concept Modeling-based Drug Repositioning

Jagadeesh Patchala[1] and Anil G Jegga[1,2,3]
[1]Department of Computer Science, [2]Division of Biomedical Informatics, [3]Department of Pediatrics, Cincinnati Children's Hospital and Medical Center, University of Cincinnati, Cincinnati, Ohio, USA

Abstract
Our hypothesis is that drugs and diseases sharing similar biomedical and genomic concepts are likely to be related, and thus repositioning opportunities can be identified by ranking drugs based on the incidence of shared similar concepts with diseases and vice versa. To test this, we constructed a probabilistic topic model based on the Unified Medical Language System (UMLS) concepts that appear in the disease and drug related abstracts in MEDLINE. The resulting probabilistic topic associations were used to measure the similarity between disease and drugs. The success of the proposed model is evaluated using a set of repositioned drugs, and comparing a drug's ranking based on its similarity to the original and new indication. We then applied the model to rare disorders and compared them to all approved drugs to facilitate "systematically serendipitous" discovery of relationships between rare diseases and existing drugs, some of which could be potential repositioning candidates.

Introduction
Drug repositioning is the process of developing new indications for existing drugs or biologics. Maximizing the indications potential and revenue from drugs that are already marketed offers a new take on the famous mantra of the Nobel Prize-winning pharmacologist, Sir James Black, "*The most fruitful basis for the discovery of a new drug is to start with an old drug*". Rational design of drug mixtures however poses formidable challenges because often the details of *in vivo* cell regulation and pathway interactions and mechanisms underlying genetic pathway regulation are obscure. Thus, several of the repositioned drugs are discovered serendipitously in the form of unexpected findings during late phase clinical studies. One of the reasons that the connection between drug candidates and their potential new indications could not be identified earlier is that the underlying mechanism associating them is either very intricate and unknown or dispersed and buried in a sea of information. Drug repositioning is predominantly dependent on two principles: i) the "promiscuous" nature of the drug and ii) targets relevant to a specific disease or pathway may also be critical for other diseases or pathways[1,2]. The latter may be represented as a shared gene or biomedical concept between a disease-disease, drug-drug, or a disease-drug. Based on this principle, some computational approaches have been developed and applied to identify drug repositioning candidates ranging from mapping gene expression profiles with drug response profiles to side-effect based similarities[3-8].

The topic model is a state-of-the-art Bayesian model for extracting semantic structure from document collections[9]. It automatically learns a set of thematic topics (lists of words or "bag of words") that describe a document collection, and assigns the topics to each of the documents in the collection with a probability value. Topic models have recently retained a lot of attention and have been used to address various issues (e.g., drug repositioning[10], word sense disambiguation in the clinical domain[11], gene-drug relationship extraction from literature[12], etc.). As a variation of classic "bag-of-words" approach, we use a "bag of concepts" approach. We first employ the UMLS Metathesaurus to identify biomedical concepts and construct a probabilistic topic model based on the concepts that appear in the disease and drug related abstracts. The resulting probabilistic topic associations are used to measure the similarity between disease and drugs and identify drug repositioning candidates (Fig. 1).

Methods
MEDLINE Abstract collection
Disease and drug-related abstracts were extracted from MEDLINE using NCBI's E-Utilities feature[13]. We created PubMed queries (using disease or drug names along with the MeSH field tag, if available) that returned respective list of articles (ranging from 100 to 10000). For topic modeling purposes, we only used PubMed search results that contained abstracts. From the collected sets of abstracts, we randomly selected 500 abstracts with mapped concepts (see section Concept Mapping) for topic modeling (Fig. 1). For validation purposes, we selected 11 disease-drug pairs representing known and candidate repositioned drugs (e.g., ropinirole-Parkinson's disease and ropinirole-Restless legs syndrome) and downloaded all the abstracts related to the disease and drug. Abstracts that cited both disease and drug are excluded from topic modeling input to avoid the over-fitting of our model to any particular drug or disease. In other words, if an abstract cites both the disease and drug from select disease-drug pairs (e.g. abstracts citing both ropinirole and Parkinson's disease), it was not used to generate the topics. As our test set, we collected the list of 1704 approved drugs from the DrugBank[14] and six rare diseases. For each of these diseases and

drugs we compiled the list of published articles and randomly selected 500 abstracts for each, at a time, for the analysis. We removed 10 drugs (from total 1704) from our drug data set because at the time of this analysis, each of these drugs had fewer than 50 publications. Our final dataset thus comprised 1694 drugs and 6 rare diseases resulting in about 850K (1700 drugs/diseases * 500 abstracts) abstracts. Our goal is to rank the 1694 drugs based on their likelihood as drug repositioning candidates for each of the selected six rare diseases as measured by their similarity to the six rare diseases. In each of the runs (total 10 runs for each disease), we changed the 500 abstracts for the rare disease and drugs and recorded the top ranked drugs.

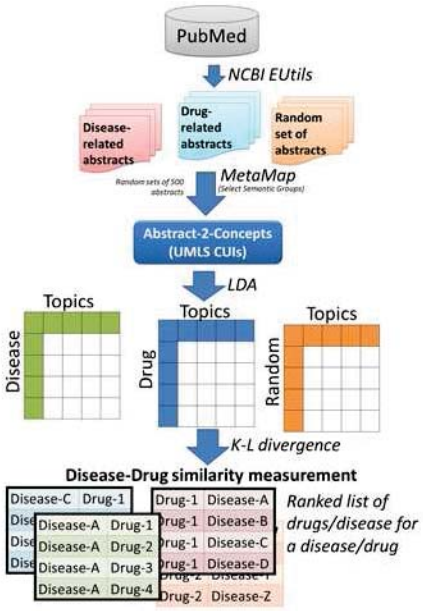

Fig. 1: Schematic representation of overall workflow. Drug and disease-related abstracts are Metamapped to generate a list of biomedical and genomic CUIs from UMLS for each drug and disease. Topic modeling is then applied followed by statistical analysis to assess the similarity between disease and drug.

Concept Mapping

To map the downloaded disease and drug related abstracts to concepts from the UMLS Metathesaurus, we used MetaMap[15] with semantic types restricted to five semantic groups, namely, Anatomy, Chemicals and Drugs, Disorders, Genes and Molecular Sequences, and Physiology. MetaMap returns the list of candidate mappings (along with their score) and all of the MetaMap identified concepts from the five select semantic groups with a score of >350 were used for topic modeling. We used the Concept Unique Identifiers (CUIs) as input instead of concept terms to avoid redundancy and increase the specificity of the model (Fig. 1).

Topic Modeling

For a document d, θ(d)= P(t) stands for the multinomial distribution over topics. Let P(w|t) be the probability distribution over words w given topic t. Then, following process generates the words in the document d, $P(w_i) = \sum_{j=1}^{t} P(\frac{Wi}{Ti} = j)P(Ti = j)$,where T is the number of topics. We used MALLET (MAchine Learning for LanguagE Toolkit)[16] a JAVA-based package to build our topic model. To determine the number of inherent topics in our data set, we first started with a topic size of 25 and calculated the log likelihood value for the trained model. The log likelihood value indicates how well the topic model fits with the data. By keeping the other parameters constant, we increased the topic size in multiples of 25 till 500 topics and calculate the log likelihood values for each scenario. The best number of topics is the one with the highest log likelihood value which was between 175-200 topics in this case. We therefore set the number of topics at 200.

Disease-drug distance assessment

We use Kullback-Leibler (KL) divergence[17] to compute the differences between the topic distributions in the selected disease and drug profiles. Given two uncertain objects P and Q and their corresponding probability distributions, KL divergence measures the similarity between two probability distributions and represents the information lost when Q is used to represent P. It is calculated as: $D_{KL}(P \parallel Q) = \sum_i P_i \log_2 (P_i/Q_i)$. Even though KL divergence is predominantly used to calculate the distance between two probability distributions, it is not a true metric as it is not symmetric. The KL divergence of P and Q is not equal to KL divergence of Q and P, unless P and Q are equal. In the current study, we therefore calculate the symmetric form of KL divergence, which is given by $D(P, Q)=D_{KL}(P \parallel Q)+D_{KL}(Q \parallel P)$. We use the intuitive idea that the drugs that are likely to be repurposed for a disease will have significantly small KL values when compared to the average of the drug distances to that disease. We capture this notion by imposing the condition that for a drug to be considered as repositioning candidate it should have a Z-score of -1.5 compared to the average. In other words, if a drug's divergence value is significantly smaller than the average divergence of all the drugs we consider it as a potential candidate for repositioning. We thus rank the drugs that have Z-score of -1.5 or lower according to their KL distance values and display them as the probable drugs that can be repositioned for that disease

Table 1: Selected examples of drugs with multiple indications

Drug	Indication-1	Indication-2
Formoterol	Asthma	Stuttering
Mitoxantrone	Multiple sclerosis	Prostate cancer
Modafinil	Narcolepsy	Bipolar disorder
Ropinirole	Parkinson's disease	Restless legs syndrome
SSRIs	Depression	Dysmorphic disorders
Terbutaline	Asthma	Preterm labor

Results

Validation

We select 6 examples of repositioned drugs (11 disease-drug pairs; 5 drugs with two indications each and one class of depression-related drugs as repositioning candidates for dysmorphology) to validate our approach (Table 1). The goal was to see how topic model-based approach will rank the drug against its multiple indications (i.e., drug-A vs. disease-1 and drug-A vs. disease-2). As described in Methods, we downloaded the drug and disease related abstracts, excluding abstracts, which cite both drug and disease. We mixed 9 random drug profiles with each disease-drug pair and calculated the rank of the original drug for the disease. We repeated this process 10 times for each disease-drug pair and calculated the accuracy, balanced accuracy, and precision as follows:

$$\text{Accuracy} = \frac{true\ positives + true\ negatives}{true\ positives + false\ positives + true\ negatives + false\ negatives}$$

$$\text{Precision} = \frac{true\ positives}{true\ positives + false\ positives}$$

$$\text{Balanced accuracy} = \frac{sensitivity + specificity}{2}$$

For the validation sets, accuracy, balanced accuracy, and precision were 0.83, 0.75, and 0.32 respectively with an AUC of 0.74. The target drug was ranked at top 61% of the time.

Table 2: Top 15 ranked drugs for six rare diseases

Rank	Polycythaemia vera	Primary myelofibrosis	Dravet syndrome	Meningioma	Narcolepsy	Netherton syndrome
1	Pipobroman	Pipobroman	Riluzole	Lomustine	Zolpidem	Clocortolone
2	Ruxolitinib	Anagrelide	Alglucosidase alfa	Temozolomide	Cloxazolam	Amcinonide
3	Tofacitinib	Ruxolitinib	Ethosuximide	Dacarbazine	Ketazolam	Prednicarbate
4	Anagrelide	Oprelvekin	Creatine	Procarbazine	Zaleplon	Alclometasone
5	Eltrombopag	Hydroxyurea	Choline	Ifosfamide	Estazolam	Flurandrenolide
6	Uracil mustard	Pomalidomide	Valproic Acid	Altretamine	Camazepam	Diflorasone
7	Oprelvekin	Tofacitinib	Lamotrigine	Carmustine	Zopiclone	Fluocinonide
8	L-Tyrosine	Pegademase bovine	Clobazam	Mechlorethamine	Halazepam	Clobetasol propionate
9	Ginseng	Bortezomib	Zonisamide	Topotecan	Quazepam	Halobetasol Propionate
10	Hydroxyurea	Busulfan	Perampanel	Dexrazoxane	Chlordiazepoxide	Flumethasone Pivalate
11	Pegademase bovine	Thalidomide	Phenacemide	Etoposide	Bromazepam	Desoximetasone
12	Pomalidomide	Becaplermin	Topiramate	Daunorubicin	Triazolam	Monobenzone
13	Acetylsalicylic acid	Lenalidomide	Pilocarpine	Vincristine	Tofisopam	Desonide
14	Bortezomib	Fludarabine	Paramethadione	Vindesine	Delorazepam	Acitretin
15	Acenocoumarol	Eltrombopag	Trimethadione	Ethiodized oil	Clotiazepam	Betamethasone

New indication search – Drug Repositioning candidates for rare diseases

For each of the six rare disorders, we identified the nearest drug neighbors by calculating the KL distance between the rare disease and all of the 1694 drugs. We repeated this 10 times by changing the profile set of the rare diseases and recorded the number of times a specific drug was ranked among top 15 out of a total ten iterations. Table 2 enlists the top 15 ranked drugs for each of the 6 rare diseases. Literature search showed that most of the top ranked drugs could be related to their mapped respective rare diseases suggesting the utility of our approach in discovering drug repositioning candidates. In the following sections we discuss a few of our findings.

In case of polycythemia vera (PV), a rare bone marrow disease that leads to an abnormal increase in the number of blood cells, the top ranked drug in our analysis is pipobroman. There are several studies reporting the efficacy of pipobroman in PV[18,19]. Ruxolitinib ranked second for PV and third for primary myelofibrosis in our analysis. Ruxolitinib has been recently reported to provide clinical benefits in patients with advanced PV[20] and in primary myelofibrosis[21].

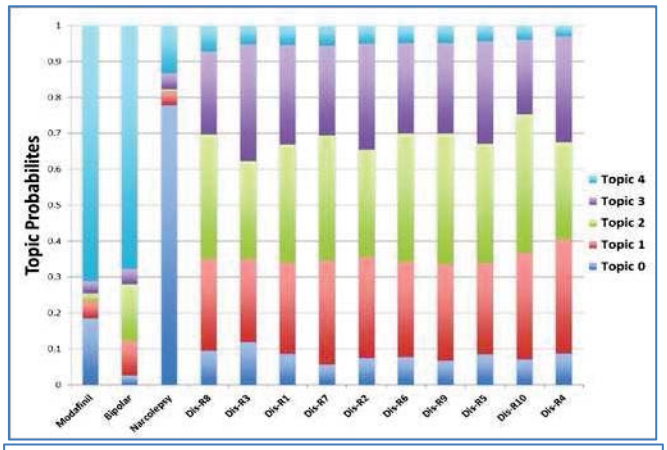

Fig. 2: Stacked bar chart showing the top five topic proportions found in modafinil (drug) and its two indications (bipolar disorder and narcolepsy) and ten random disease sets.

Dravet syndrome, a rare genetic epileptic encephalopathy, is primarily caused by mutations in the voltage-gated sodium channel *SCN1A* gene. The top ranked drug in our analysis for Dravet syndrome is riluzole, a sodium channel inhibitor. Although loss-of-function mutations are common in Dravet syndrome, a gain-of-function mutation[22] and duplications[23] in *SCN1A* have also been reported suggesting that sodium channel inhibitors like riluzole may be useful in such cases. Interestingly, riluzole was first developed as an anti-epileptic drug but is now used for treatment of amyotrophic lateral sclerosis[24]. The other top

ranked drugs for Dravet syndrome were various antiepileptic drugs. Likewise, for meningiomas, a diverse set of tumors arising from the meninges, among the top ranked drugs were several candidates that are currently investigated for various forms of brain tumors.

We also note that a high conceptual similarity between a disease and drug may not always suggest alternate indication but semantic relatedness or potential contraindication or even drug related side-effects. For example, in narcolepsy, a rare sleep disorder that causes excessive sleepiness and frequent daytime sleep attacks, all of the top ranked drugs are drugs used in the management of insomnia. This implies that although conceptually related, the top ranked drugs are not recommended for use in narcolepsy. Likewise, in case of Netherton syndrome, a rare and severe, autosomal recessive form of ichthyosis associated with mutations in the *SPINK5* gene and currently with no known cure, the top ranked drugs in our analysis were mostly from the drug class corticosteroids. However, in practice, while topical corticosteroids may be helpful in older children, they are not usually recommended in infants as impaired barrier function in Netherton's syndrome can lead to increased cutaneous absorption resulting in complications such as pituitary adrenal axis suppression[25].

Topic Concepts as indicators for repositioning

Topic (Topic 4 – Fig. 2) shared between modafinil and biopolar disorder showed words/concepts related to neuropsychiatric or behavioral conditions (e.g., *mental Depression, major depressive disorder, attention deficit hyperactivity disorder, antidepressive agents, mental association, methylphenidate, attention, lithium, sleep*, etc.) while topic 0 shared between modafinil and narcolepsy was predominantly sleep-related (*sleep disorders, cataplexy, narcolepsy-cataplexy syndrome, sleep, REM, obstructive sleep apnea, drowsiness, hypersomnia, wakefulness, REM sleep behavior disorder*, etc.).

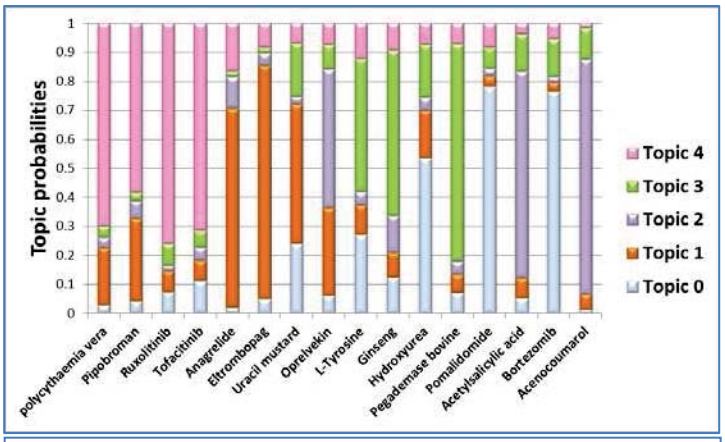

Fig. 3: Stacked bar chart showing the top ranked drugs for PV along with distribution of top five topic proportions

In case of the rare disease PV, topic 4 (Fig. 3) which was shared between the three top ranked drugs (pipobroman, ruxolitinib and tofacitinib) and PV are related to etiology of PV (e.g., *Janus kinase 2, Primary Myelofibrosis, Myeloproliferative disease, Essential Thrombocythemia, Alleles, Signal Transduction, cytokine, Janus kinase, Thrombosis, Janus kinase 1, Janus kinase 3, Interleukin-6,* etc.). Drugs ranked 4th an 5th (anagrelide and eltrombopag), interestingly are, used for thrombocytosis and thrombocytopenia and are related to PV through topic 1 (*Blood Platelets, Essential Thrombocythemia, Thrombocytopenia, Thrombocytosis, Megakaryocytes, Idiopathic Thrombocytopenic Purpura, Thrombopoiesis, Thrombosis, Thrombopoietin, Thrombus,* etc.) representing platelets and platelet-related concepts. Anagrelide reduces the platelet count and is reported to be beneficial in some patients and is recommended as second line-therapy in PV[26].

Discussion

We used topic modeling to estimate the probability distribution of topics for each of the drugs or diseases and assess the disease-drug similarity. While more extensive validation studies are required to further validate our approach, results from our preliminary validation tests and rare diseases demonstrate the utility of our approach. While the accuracy of our approach is high, the lower precision rate may be partially due to the small size of the validation sets. The novelty of our approach is several fold: first, instead of using the abstracts directly, we use mapped biomedical concepts for topic modeling which would increase the specificity and also overcome the problem of biomedical stop words to some extent; second, apart from using UMLS CUIs for topic modeling, we filter the CUIs further limiting only those belonging to relevant semantic groups; third, our approach compares disease and drug directly unlike previous approaches which focus on either drug-drug or disease-disease relationships to find drug repositioning candidates. Further, to the best of our knowledge, our study is the first to use topic modeling on MEDLINE abstracts for drug repositioning candidate discovery for rare diseases.

Some of the planned extensions for the current model relate to methodology and the data sets used. For instance, in the current study, based on the log likelihood value, we selected 200 as the topic size. However, we plan to investigate different methods and metrics for judging the optimal number of topics more systematically. While we

focused on the UMLS concepts for mapping biomedical and genomic concepts, UMLS has certain limitations especially with gene and genomic annotation representation in the UMLS Metathesaurus. We plan to supplement this by including additional resources for gene and genomic annotations (e.g., other biomedical ontologies via NCBO Annotator[27]). Since, literature related to drugs and diseases are constantly updated, the dynamic and temporal nature of the disease and drug concepts can be utilized for a more robust drug repositioning and computational pharmacovigilance systems. Although we focus on drug repositioning in this study, based on our results, the current approach can also be employed to understand the molecular basis of side-effects or suggest safer alternatives (e.g., drugs with fewer side-effects) by ranking drugs against diseases based on side-effects topics. Lastly, as a future extension, we plan to compare all of the rare disorders to approved drugs using the current approach.

References

1. Pujol A, Mosca R, Farres J, Aloy P. Unveiling the role of network and systems biology in drug discovery. Trends Pharmacol Sci. 2010;31(3):115-23.
2. Sardana D, Zhu C, Zhang M, Gudivada RC, Yang L, Jegga AG. Drug repositioning for orphan diseases. Brief Bioinform. 2011;12(4):346-56.
3. Campillos M, Kuhn M, Gavin AC, Jensen LJ, Bork P. Drug target identification using side-effect similarity. Science. 2008;321(5886):263-6.
4. Hu G, Agarwal P. Human Disease-Drug Network Based on Genomic Expression Profiles. PLoS ONE. 2009;4(8):e6536.
5. Hurle MR, Yang L, Xie Q, Rajpal DK, Sanseau P, Agarwal P. Computational drug repositioning: from data to therapeutics. Clin Pharmacol Ther. 2013;93(4):335-41.
6. Iorio F, Bosotti R, Scacheri E, et al. Discovery of drug mode of action and drug repositioning from transcriptional responses. Proc Natl Acad Sci U S A. 2010;107(33):14621-6.
7. Lamb J, Crawford ED, Peck D, et al. The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease. Science. 2006;313(5795):1929-35.
8. Sirota M, Dudley JT, Kim J, et al. Discovery and preclinical validation of drug indications using compendia of public gene expression data. Sci Transl Med. 2011;3(96):96ra77.
9. Blei DM, Ng, A.Y., Jordan, M.I. Latent Dirichlet Allocation. Journal of Machine Learning Research. 2003;3:993-1022.
10. Bisgin H, Liu Z, Kelly R, Fang H, Xu X, Tong W. Investigating drug repositioning opportunities in FDA drug labels through topic modeling. BMC bioinformatics. 2012;13 Suppl 15:S6.
11. Chasin R, Rumshisky A, Uzuner O, Szolovits P. Word sense disambiguation in the clinical domain: a comparison of knowledge-rich and knowledge-poor unsupervised methods. Journal of the American Medical Informatics Association : JAMIA. 2014;21(5):842-9.
12. Wu Y, Liu M, Zheng WJ, Zhao Z, Xu H. Ranking gene-drug relationships in biomedical literature using Latent Dirichlet Allocation. Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing. 2012:422-33.
13. Sayers E. E-utilities quick start 2008. Available from: http://www.ncbi.nlm.nih.gov/books/NBK25497/.
14. Knox C, Law V, Jewison T, et al. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. Nucleic acids research. 2011;39(Database issue):D1035-41.
15. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proceedings / AMIA Annual Symposium AMIA Symposium. 2001:17-21.
16. McCallum AK. MALLET: A Machine Learning for Language Toolkit 2002. Available from: http://mallet.cs.umass.edu/.
17. Kullback S. Information theory and statistics. New York: John Wiley and Sons; 1959.
18. Kiladjian JJ, Chevret S, Dosquet C, Chomienne C, Rain JD. Treatment of polycythemia vera with hydroxyurea and pipobroman: final results of a randomized trial initiated in 1980. J Clin Oncol. 2011;29(29):3907-13.
19. Passamonti F, Brusamolino E, Lazzarino M, et al. Efficacy of pipobroman in the treatment of polycythemia vera: long-term results in 163 patients. Haematologica. 2000;85(10):1011-8.
20. Verstovsek S, Passamonti F, Rambaldi A, et al. A phase 2 study of ruxolitinib, an oral JAK1 and JAK2 Inhibitor, in patients with advanced polycythemia vera who are refractory or intolerant to hydroxyurea. Cancer. 2014;120(4):513-20.
21. Harrison C, Kiladjian JJ, Al-Ali HK, et al. JAK inhibition with ruxolitinib versus best available therapy for myelofibrosis. N Engl J Med. 2012;366(9):787-98.
22. Volkers L, Kahlig KM, Verbeek NE, et al. Nav 1.1 dysfunction in genetic epilepsy with febrile seizures-plus or Dravet syndrome. Eur J Neurosci. 2011;34(8):1268-75.
23. Marini C, Scheffer IE, Nabbout R, et al. SCN1A duplications and deletions detected in Dravet syndrome: implications for molecular diagnosis. Epilepsia. 2009;50(7):1670-8.
24. Eijkelkamp N, Linley JE, Baker MD, et al. Neurological perspectives on voltage-gated sodium channels. Brain. 2012;135(Pt 9):2585-612.
25. Eichenfield LF, Tom WL, Berger TG, et al. Guidelines of care for the management of atopic dermatitis: section 2. Management and treatment of atopic dermatitis with topical therapies. J Am Acad Dermatol. 2014;71(1):116-32.
26. Finazzi G, Barbui T. Evidence and expertise in the management of polycythemia vera and essential thrombocythemia. Leukemia. 2008;22(8):1494-502.
27. Shah NH, Bhatia N, Jonquet C, Rubin D, Chiang AP, Musen MA. Comparison of concept recognizers for building the Open Biomedical Annotator. BMC bioinformatics. 2009;10 Suppl 9:S14.