Crasto et al., "GenDrux: A supervised biomedical literature searching tool that recognizes gene expression-based drug sensitivity in breast cancer"

This paper presents an automatically collected repository of gene-drug-disease relations extracted from PubMed abstracts, currently focused on breast and other cancers. The authors use resources such as GeneOntology annotations, NCBI's Entrez Gene, etc. to map PubMed abstracts to genes that are discussed therein, achieving high data quality. The database can be accessed publicly via their webserver. Searches lead to evidence in the literature for gene/drug relations in the context of breast cancer; genes and drug are linked to external resources for further information (Amigo and drugs.com).

- The suggestions I had in the previous review round were addressed by the authors, either in the revised manuscript or the cover letter, where appropriate.

Major revisions

- Due to the methodology for searching and filtering (relying on existing annotations), the database seems outdated (GeneRIFs will only be extracted by NCBI staff etc. a certain time after an article has been published). The latest articles seem to be from 2008, at least in the online version, although the authors report a large portion coming from 2009 and 2010. I tried searches for Brca1 and -2, estrogen, cisplatin, estradiol, and breast cancer, showing no results for 2009/10.

- Please define your usage of the terms "supervised" and "unsupervised" more clearly. In traditional machine learning / AI, which here would include NER etc., "supervised" refers to methods that rely on a manually created resource for learning a model (such as a set of texts annotated with gene/drug/disease names by hand). Abner, Banner, and Lingpipe are supervised methods in that sense, although you describe them as unsupervised (p.4). "Supervised" as you use it refers more to manually created resources, not to methods.

- Similarly, you describe GenDrux as "supervised" b/c you take your mapping of PubMed abstract to, for instance, genes from gene annotations in NCBI Entrez Gene, GEO, etc. [This step should be described more clearly in any case]. However, this methodology leads to the same problem of low recall that you use
as an argument to discard methods such as Abner and Banner to find genes automatically (p.4: "subject to issues of precision and recall"), as GO, EntrezGene, etc. never have all PubMed records listed that are relevant for a particular gene (or vice versa). You might thus miss a lot of abstracts through your filtering, if no gene was annotated for a record (yet).

- As the search and filtering in your method are important steps, they should be explained in more detail were the assignment of genes/drugs/diseases (disease=breast cancer or other cancers) is concerned. Please list all resources that you use for mapping PubMed IDs to GeneIDs etc. How do you finally "spot" each gene name in a PubMed record once you know it is there via the mapping? The gene annotation files contain only IDs (or some synonyms), but the authors of a particular paper might have used a different alias for that gene altogether. Will you miss the gene's mention in these cases? Will that paper then be filtered out, or do you still keep it, but the gene would not be highlighted in your tool (b/c you don't know at which position(s) in the text it occurs)?

Minor essential revisions

- It would be very useful to be able to sort results of the online tool by publication date, PubMed ID, gene name / drug name.

- In addition to explaining how you find the (positions of) genes in each abstract, please describe how you find drugs in each abstract, as well as diseases other than breast cancer (other types of cancer).

- As mentioned above, your recall is also affected by articles from 2009/10 not showing up in the online tool.

Discretionary revisions

- change "name entity recognition" to "named entity recognition" throughout

- Why did you pick Amigo and especially drugs.com to link genes and drugs to external sources? In my opinion, these are too high-level. There are more resources out there that provide specific information that are interesting to clinicians, biologists, biochemists, etc. (DrugBank, PubChem, PharmGKB; EntrezGene, UniProt, HUGO, GeneCards; etc.) -- especially since you want to help "a physician make an informed decision about prescriptions and diagnoses after assessing the impact of a drug on patients' genetic profile", such data should only be one or two clicks away.

Level of interest: An article whose findings are important to those with closely related research interests

Quality of written English: Needs some language corrections before being published
Statistical review: No, the manuscript does not need to be seen by a statistician.

Declaration of competing interests:
I declare that I have no competing interests.