Using Semantic Web Technologies to Reproduce a Pharmacovigilance Case Study

Michiel Hildebrand[1,2], Rinke Hoekstra[1] and Jacco van Ossenbruggen[1,2]

[1] VU University Amsterdam,The Netherlands
[2] Centrum Wiskunde & Informatica, Amsterdam, The Netherlands

Abstract. We provide a detailed report of a reproduction study of a paper published in the International Journal of Medical Sciences (IJMS). We first use the PROV-O ontology to model our reconstruction of the computational workflow of the original experiment and to systematically explicate all information that is needed for an reproduction study. We then identify which part of the required information is published in the IJMS paper and what part is missing. We then discuss our reproduction of this workflow, following the original as much as possible. Again, we use PROV-O to precisely define our version of the workflow, including our version of the information that was missing in the IJMS paper of the study. Finally, we generalize from the specific cased described in the original paper by providing a web service that allows mining for arbitrary drug-adverse event pairs.

1 Introduction

Reproducing scientific results is often more an art than science. By describing a concrete case study we show how we used PROV-O to systematically analyse a paper from a different field, written by authors we do not personally know. We attempt to reconstruct the provenance graph of the original experiment by carefully studying the description of the method, the statistics and the results provided either directly in the paper or other sources that the paper refers to. We formalized our reconstruction using the PROV-O ontology. The formalization makes the dependencies between the intermediate steps explicit, which should allow us to systematically investigate how the results presented in the paper were computed. To reproduce the results we need to understand the input and output behavior of the computations modeled by the `prov:Activity` nodes. The properties of the input and output `prov:Entity` can help to verify wether this understanding is correct.

The paper we selected is the Open Access article *Adverse Event Profiles of 5-Fluorouracil and Capecitabine: Data Mining of the Public Version of the FDA Adverse Event Reporting System, AERS, and Reproducibility of Clinical Observations* published in the International Journal of Medical Sciences (IJMS) [12]. The paper describes a computational data mining study on public data and appears to be a good candidate for a reproduction study. We use this paper as a case study to provide insights in the problem of reproducing scientific results, we do not aim to criticize this particular paper in any way.

The topic of the paper is an example of *pharmacovigilance* which is defined by the World Health Organization as "the science and activities relating to the detection, assessment, understanding and prevention of adverse effects or any other medicine-related problem"[3]. Computational studies play an important role in Pharmacovigilance to detect drug side-effects. Such studies are an economic way to generate hypotheses before performing costly clinical reviewing [13]. The IJMS paper [12] follows a typical scenario in pharmacovigilance: the use of a database with reports of adverse events (AE) to find disproportional correlations between a drug and an adverse reaction. In this example, the Adverse Event Reporting System of the US Food and Drug Administration (FAERS) is used to compare adverse effects of drugs.

While the FAERS database itself is publicly available, it is not trivial to reproduce the results of the experiments that use this database. Results and tools are described in scientific publications, but tools and (intermediate) results are typically not available. Our case study demonstrates in detail what prevents reproduction. From the observations of this study we derive initial requirements to support studies of drug side-effects that can be fully reproduced.

2 Related work

This section gives a brief overview of related work on data publication, scientific workflows and provenance.

The requirement for reproducibility [14] has been a key motivator for an increased interest in data sharing and publication, especially in fields dealing traditionally with ever growing datasets, e.g. [1]. Even though data sharing does not always immediately benefit the individual researcher, the potential for the scientific community is significant [15]. Funding agencies, keen on maximizing impact and reducing fraud, are now actively requiring data sharing. For example, both the US National Science Foundation and the EU now require data management plans for all proposals they consider.[4] Note that also in areas that focus on human action, such as in human computer interaction, replication has part of the research agenda[5].

However, as becomes clear in this paper as well, raw data publication (such as FAERS) is in itself not sufficient for reproducible research. Data often needs to be moulded and transformed to a new data model before it becomes suitable for answering a particular research question. This data preparation step can take between 60 to 80% of data-oriented research tasks [6]. Workflow systems [4], provide mechanisms for reproducing scientific conclusions, based on shared data.

[3] http://www.who.int/medicines/areas/quality_safety/safety_efficacy/pharmvigi/en/

[4] See http://www.nsf.gov/bfa/dias/policy/dmp.jsp and http://europa.eu/rapid/press-release_SPEECH-13-236_en.htm

[5] http://www.cs.nott.ac.uk/~mlw/replichi.php

The benefit for individual researchers publishing a workflow, is that workflows are *executable* procedures that can be run against various inputs. Workflows can be shared and reused through social platforms such as myExperiment[6]. Curated workflow descriptions [8] combined with original data, can serve as self-contained *research objects* [3].

There are, however, two drawbacks to using a workflow system. Firstly, workflow descriptions are inevitably tied to the system used, and thus constrained to the types of operations supported by the system. Secondly, not all steps of interest in a scientific research process are necessarily of a *computational* nature, e.g. consider the information conveyed through the reuse of texts in scientific discourse. Though in its early stages, work on automatic *provenance reconstruction* [10] is a promising approach to making explicit the temporal and causal dependencies between individual elements of scientific output.

The overarching requirement for reproducible research is an explicit account of what processes and activities led from original input, albeit data, texts, other media, to the contribution of a scientific publication. The PROV standard of the W3C [11], based on ten earlier provenance models, such as the Open Provenance Model[7] and the Provenance Vocabulary[8], provides a standard vocabulary and semantics for expressing plans (workflows), process execution, dependencies between entities and processes, and agent involvement. The PROV-O ontology is a vocabulary for expressing PROV as Linked Data.[9] Most scientific workflow systems allow provenance tracking of workflow execution, and allow exporting it to PROV or a compatible format. The consumption of provenance information by applications is gradually receiving more attention [9]. The ProvBench repository[10] has the objective to bootstrap the development of systems for the visualization, analysis and understanding of provenance graphs.

3 Basic concepts in Pharmacovigilance

Various organizations maintain reporting systems of adverse events. The World Health Organization (WHO) maintains vigiBase, the US Food and Drug Administration (FDA) maintains the Adverse Event Reporting System (FAERS) and many countries maintain their own system. These organisations provide functionality for medical professionals to submit reports of adverse events that they encountered with their patients. A report in the FAERS database contains a list of the medication that the patient received and a list of adverse events. In addition it may contain information about the patient such as the gender and age. Unique of the the FAERS database is that it is publicly available on the Web. XML and CSV files for every yearly quarter starting at 2004 are available for download.

[6] See `http://www.myexperiment.org`

[7] See `http://purl.org/net/opmv/ns`

[8] See `http://purl.org/net/provenance/ns`

[9] See `http://www.w3.org/TR/prov-o/`.

[10] `https://sites.google.com/site/provbench/`

Adverse event databases are used in pharmacovigelance research to detect side effects of drugs. An important part of this research focusses on the detection of side effects of new drugs that appear on the market. The WHO has an extensive program for this research[3], and involves large scale data mining of adverse event databases. Other research focusses on the side effects of sets of specific drugs. These studies are typically motivated by clinical evidence.

Both types of research depend on methods to detect a disproportional correlation between a drug and an associated adverse event. The most common methods are the proportional reporting ratio (PRR), the reporting odds ratio (ROR), the information component (IC) and the empirical Bayes geometric mean (EBGM). All are based on the expected frequency relative to all drug event pairs that are available in the database. Calculating signals with these methods requires a 2x2 contingency table, as shown in Table 1. This table contains (a) the number of mentions of a drug together with a mention of a reaction (an adverse event), (b) the number of mentions of all other drugs and that reaction, (c) the number of mentions of the drug and all other reactions and (d) the number of mentions of all other drugs and all other reactions. According to [5] the PRR is calculated from this table using Eq. 1.

$$PRR = \frac{a/(a+c)}{b/(b+d)} \tag{1}$$

The expected value for a PRR is one and values above it indicate the strength of the association. In addition, the strength of a statistical association can be calculated using a standard chi-squared test.

$$\chi^2 = \frac{(ad-bc)^2(a+b+c+d)}{(a+b)(c+d)(b+d)(a+c)} \tag{2}$$

According to [5] a signal is detected between a drug and an adverse event if the PRR is at least 2, the chi-squared is at least 4 and there are at least 3 or more cases mentioning the drug and the event. We refer to the literature for the details of ROR [16], IC [2] and EBGM [17]. A comparison of these methods is reported in [18].

To compute the 2x2 contingency table one needs to collect all mentions of a particular drug and a particular adverse event. Collecting the adverse event mentions is straightforward because in the FAERS database they are consistently identified with the preferred terms from the Medical Dictionary for Regulatory Activities[11] (MedDRA). The drug names in the FAERS database are, however, not standardized. The same drug may be entered in to the database in various forms. For example, drug names are entered with or without dosage information, method (e.g. oral, injection) and other additions. Some have entered the drug name, while others used the brand or trade name and again others the active ingredient. There are various spelling variations and synonyms. To properly fill the 2x2 contingency table one has to deal with the variations in drug names.

[11] `http://www.meddramsso.com/`

	Drug of interest	All other drugs	
Reaction of interest	a	b	a+b
All other reactions	c	d	c+d
	a+c	b+d	a+b+c+d

Table 1. 2x2 contingency table to calculate disproportionality measurements (adapted from [5]).

4 Case study

Our target IJMS paper [12] investigates the so called safety profiles of two types of drugs that are used in the treatment of cancer. The first drug is 5-Fluorouracil (5-FU), which was traditionally used for the treatment of solid tumors. This drug was given by injection or infusion. Due to the high risks and costs of this type of treatment the pharmaceutical industry developed a class of drugs known as oral fluoropyrimidines, from which Capecitabine is the most well known one. Clinical trials that compared the use of Capecitabine against 5-FU favor the use of the first. Due to limitations of the clinical trials the picture is, however, not complete. For example, the trials do not provide evidence for adverse events that occur at relative low frequencies. The aim of the paper is to test the conclusions drawn from the trials and provide additional evidence for lower frequency adverse events.

In the IJMS paper the authors describe the method to detect the signals for Capecitabine and 5-FU with various adverse events. As a first step towards reproduction of this study we formalized the steps and their dependencies using the PROV-O ontology. In addition, we describe the information provided in the paper that could help in the reproduction.

4.1 Provenance reconstruction

Figure 1 in Appendix A shows a reconstruction of the provenance graph for the computation of the PRR for 5-FU with the adverse event Leukopenia and the PRR of Capecitabine with Leukopenia.

Original FDA datasources The workflow starts at the bottom with datasources obtained from the FDA. The website of the FDA contains ZIP files for each yearly quarter[12]. For each quarter there are two versions available, ASCII and SGML. The former contains a dump of the database in the form of 7 CSV files, while the latter contains a single SGML file. The authors of the IJMS paper used the ASCII versions for the first quarter of 2004 up to the last quarter of 2009, a

[12] `http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/ucm083765.htm`

total of 24 files. In the provenance graph the quarterly files are represented by individual nodes, but for the sake of clarity we do not show all nodes. The ZIP files from the FDA contain a document that describes the structure of the CSV files and instructions how to interpret them.

Report aggregation The paper mentions that the total dataset contains 2,231,029 reports. From this we conclude that an aggregation step was performed. In the provenance graph the aggregated dataset is represented by the node with the label `A.FAERS`. This aggregated dataset is the starting point for two cleanup activities. First, the authors removed superfluous reports as the data contains updated versions of a report as separate records. The paper refers to the documentation from the FDA in which it is advised to keep only the most recent report for a specific case. The resulting dataset is labeled `B.FAERS` in the provenance graph, and contains (according to the paper) 1,644,220 reports.

Drug name normalization In the second cleanup step the drug names are normalized: *all drug names were unified into generic names by a text-mining approach.* The paper does not provide details of this text-mining approach. The paper does explain that the cleanup includes the correction of spelling errors. For this purpose GNU Aspell is used to detect spelling errors and the suggested corrections are manually confirmed by *working pharmacists.* It is unclear how many spelling corrections were made. Finally, *foods, beverages, treatments (e.g. X-ray radiation), and unspecified names (e.g. beta-blockers)* were removed. It is unclear from the paper if this removal step is manual or automatic. The result of the normalization activity is represented in the provenance graph by the node `C.FAERS`.

Co-occurrence selection The paper mentions that after the drug name normalization the dataset contains 22,017,956 co-occurrences of drugs and events. A drugname and an adverse event co-occur if they are mentioned together in a report. The activity of counting co-occurrences is modeled as an explicit step and the output is the node with label `D.co-occurrences`.

Contingency table To compute the PRR values from the set of co-occurrences a 2x2 contingency table is required for each drug-adverse event pair. Populating the table requires the selection of the required subsets of co-occurrences. The graph contains activities to create the tables for 5-FU with Leukopenia and Capecitabine with Leukopenia. The resulting tables are shown as the nodes `5FU-Leukopenia` and `Capecitabine-Leukopenia`. The IMJS paper does not explicitly contain the 2x2 table for any of the drug-adverse event pairs, but using the values mentioned in the paper we can partially reconstruct the tables, see Table 2. In this table the values in bold font come from the paper, the italic ones can be trivially calculated from these. The question marks represent values which we will try to reverse engineer in the next section.

	5-FU	All other drugs	
Leukopenia	**277**	?	?
All other reactions	*40,007*	?	?
	40,284	*21,977,672*	**22,017,956**

	Capecitabine	All other drugs	
Leukopenia	**115**	?	?
All other reactions	*34,813*	?	?
	34,928	*21,983,028*	**22,017,956**

Table 2. Partial 2x2 contingency tables for 5-FU - Leukopenia and Capecitabine - Leukopenia from the numbers provided in the IJMS paper. The numbers in italic are calculated from the numbers that are given in the paper.

PRR values The final PRR values and the results of the chi-squared test are provided in the IMJS paper. In the provenance graph they are represented as the end nodes, e.g. `PRR 5FU-Leukopenia`. Note that to recalculate the values for the PRR and chi-squared tests we need to obtain the missing values in Table 2.

5 Reproduction experiment

We first tried to recalculate the missing numbers in the 2x2 contingency tables using the information given in the paper. Next we tried to reproduce the subsets of drug-adverse event pairs that underly the 2x2 co-occurrences using the original FAERS data from the FDA website, and thus reproducing the entire workflow. Further details of the reproduction are available at the Website accompanying this paper `http://www.few.vu.nl/~michielh/lisc2013/`.

5.1 Missing numbers and formulas

Using the PRR values given in the paper and the PRR formula cited by the paper, we should be able reconstruct the missing values from the 2x2 contingency tables. Note that while we do not know values for b, the number of mentions of an adverse event in co-occurrence with all other drugs, we do know the values for $(b + d)$. Based on Eq. 1, we should thus be able to calculate the values for b by using Eq. 3.

$$b = \frac{a/(a + c)}{PRR} \times (b + d) \tag{3}$$

Knowing b, we should also be able to compute the total number of mentions of an adverse event in the database $a + b$. For example, using the PRR value

	5-FU	All other drugs	
Leukopenia	277	28,585	28,862
All other reactions	*40,007*	21,949,087	21,989,094
	40,284	*21,977,672*	22,017,956

	Capecitabine	All other drugs	
Leukopenia	115	28,747	28,862
All other reactions	*34,813*	21,954,281	21,989,094
	34,928	*21,983,028*	22,017,956

Table 3. Reproduction of 2x2 contingency tables for 5-FU - Leukopenia and Capecitabine - Leukopenia.

for 5-FU (5.282) and the numbers from the partial contingency table, Table 2, the total number of mentions of Leukopenia should be 28,887. Surprisingly, this number is different when calculated from the PRR for Capecitabine (2.520), namely 28,952. For the other adverse events mentioned in the paper we also found a difference when calculated with the PRR of 5-FU or with the PRR of Capecitabine. These differences are all bigger than can be explained by rounding errors. After more in-depth literature study we discovered that different formulas are used to calculate the PRR. For example, the IJMS paper also cites [7] that uses the formula given in Eq. 4:

$$PRR_2 = \frac{a/(a+c)}{(a+b)/(a+b+c+d)} \tag{4}$$

Unfortunately, we do not get a constant number for $a + b$ with this formula either. However, after some experimentation we discovered that with Eq. 5 we achieve a constant number for the mentions of Leukopenia, 28,862. Also for the other adverse events this formula results in a constant number. From this we conclude that while Eq. 5 is given nor cited by the IMJS paper, it is most likely the formula used to calculate all PRR values mentioned in the paper (!).

$$PRR_3 = \frac{a/c}{(a+b)/(a+b+c+d)} \tag{5}$$

Now that the total number of mentions of Leukopenia is known (a+b) we can complete the 2x2 contingency tables, see Table 3. Using this table it is also possible to, modulo rounding errors, successfully reproduce the values from the chi-squared tests with Eq. 2. Now we know how to compute the basis statistics reported by the paper, we can try to reproduce the entire experiment.

5.2 Workflow reproduction

As it is unclear how the drug name normalization was performed, we decided not to reproduce this on the entire dataset. We focus on the two drugs mentioned in the IMJS paper: 5-FU and Capecitabine. Our aim is to approximate the PRR values for these drugs and Leukopenia. The provenance graph of our reproduction is available at `http://www.few.vu.nl/~michielh/lisc2013/prov/`. We encourage the reader to access this graph. The `prov:Entity` nodes in this graph are clickable and point to the underlying data. In this way we provide access to the intermediate datasets, which is an essential ingredient to successful reproduction of computational workflows. Currently, we are investigating normalization of all drug names in the FAERS dataset.

Original FDA datasources Similarly as the study reported in the IMJS paper we downloaded the 24 quarterly dumps (from the beginning of 2004 to the end of 2009) from the FDA website.

Report aggregation by conversion to RDF We choose to aggregate the quarterly files into a single dataset by first converting them to RDF and then storing these in a triple store. The total number of reports in our RDF dataset is 2,231,038, this is 9 reports more than reported in the IMJS paper. It is unclear where the difference comes from. We can, however, confirm that the conversion to RDF did not alter the original reports, as the original CSV files combined also contain 2,231,038 unique report identifiers[13]. The conversion from CSV to RDF was performed using SWI-Prolog and the RDF conversion toolset[14]. Details of the conversion, the resulting RDF and the SPARQL endpoint are available at `http://www.few.vu.nl/~michielh/lisc2013`.

Duplicate removal The duplicate removal step was performed on the RDF dataset. We first grouped all reports with the same case number and for each group selected the report with the highest report identifier. We removed the other reports from the database. The resulting dataset contains 1,664,078 reports, this is 142 less than reported in the IMJS paper. We can't explain this difference.

Drug name normalization Instead of normalizing all the drug names, we tried to find all the mentions for our drugs of interest: 5-FU and Capecitabine. We explored four methods to find different mentions for these drug names.

1. We selected the mentions that contain the drug name itself. For Capecitabine this returns many mentions of *capecitabine*, but also many variations such as *capecitabine tablet 1000 mg*, *capecitabine roche laboratories inc* and *capecitabine 2000 mg po as divided doses daily*. In total we find 337 different mentions containing Capecitabine.

[13] The total number of unique report identifiers in the CSV files from the FDA is computed with a unix bash script: `cut -d$ -f1 DEMO0*.TXT | sort -u | wc -l`

[14] `http://semanticweb.cs.vu.nl/xmlrdf/`

2. We selected mentions of a brand name associated with the drug. For example, Capecitabine is sold under the brand name *Xeloda*. To get the brand names for a drug we used the Open Data Drug & Drug target Database, Drugbank[15]. For Capecitabine we found in Drugbank one brand name (*Xeloda*). Using this brand name 14 mentions are found in the FAERS dataset. From these mentions 4 already contain Capecitabine, e.g. *xeloda capecitabine*. For 5-FU Drugbank contains 25 brand names, such as *Adrucil* and *Fluoroplex*. For 6 of these 25 brand names, additional mentions were found in the FAERS dataset.
3. We used Drugbank to find synonyms associated with a drug name. For example, Capecitabine is also known as R340. However, no mentions of this synonym are found in the FAERS dataset. For 5-FU no synonyms are found in Drugbank.
4. We selected mentions of drug names that were spelled differently. Similarly as was reported in the IMJS paper we used GNU Aspell. Aspell contains dictionaries for many languages, but these are not very useful for drug names. Therefore, we created our own Aspell dictionary using the drug names mentioned in Drugbank. With this dictionary we generated spelling variations for all drug mentions in the FAERS dataset. For each drug mention we added the highest ranked suggestion as an alternative label to the database. For example, *capecitabine* was suggested for the drug mention *capecitabin* and *capecitapine*. When using these spelling suggestions we could retrieve for Capecitabine another 30 different mentions. From these 10 already contained the correct spelling, e.g. *capeciabine capecitabine*.

Co-occurrence selection Without drug name normalization our dataset contains a total of 23,865,029 drug-adverse event co-occurrences, 1,847,073 more than reported in the IMJS paper. This larger number of co-occurrences can be explained by the fact that we did not remove *foods, beverages, treatments (e.g. X-ray radiation), and unspecified names (e.g. beta-blockers)*, as was mentioned in the IMJS paper. In addition, drug names for a single report may contain multiple treatments each containing a different drug mention. For example, a report may contain treatment with the mention *capecitabine 500 MG* and another with the mention *capecitabine 1000 MG*. In other words, the patient received two treatments, and in the second treatment the dosage of Capecitabine was increased. Without drug name normalization these mentions are counted as two co-occurrences, whereas after normalization they will be counted as a single co-occurrence. Considering the formula for PRR in Eq. 5 this difference is reflected in the denominator, the total number of co-occurrences ($a+b+c+d$) as well as the total number of co-occurrences with a specific adverse event ($a+b$).

Contingency table Using the four methods to find drug mentions we selected the set of co-occurrences corresponding to the cells of the 2x2 contingency. The total number of co-occurrences with Leukopenia (a+b) that we found is 30,724.

[15] `http://www.drugbank.ca/`

A difference of 1,862 with the number reported in the IMJS paper. This can also be explained by the lack of drug name normalization. The total number of co-occurrences with 5-FU is 42,115, 1831 more than reported in the IMJS paper. For Capecitabine 37,973 co-occurrences are found, 3045 more than in the IMJS paper. We conclude that the four drug name selection methods find more mentions of the two drugs. Currently we are investigating if and why drug mentions are falsely included. We found 289 co-occurrences for 5-FU with Leukopenia. This is 12 more than reported in the IMJS paper. For Capecitabine 122 co-occurrences were found with Leukopenia, 7 more than the 115 reported in the IMJS paper.

PRR values Using the values in the reproduced 2x2 contingency tables and Eq. 5 the PRR for 5-FU with Leukopenia is 5.367 compared to 5.282. The chi-squared test is 1019.763 compared to 952.334. For Capecitabine with Leukopenia the PRR is 2.503 compared to 2.520 in the IMJS paper, and the chi-squared test is 109.661 compared to 103.730.

6 Discussion

Reproducing the study described in the IMJS paper required substantial effort, it was difficult to verify the results of the intermediate datasets and almost impossible to analyze the differences in the reproduction. And this is all despite the fact that the IMJS paper of the case study at first sight clearly describes the method and results. By formalizing the computational workflow in PROV-O it became possible to systematically investigate the intermediate steps. We believe that sharing such provenance graphs is a first step in simplifying the reproduction of computational workflows. The next step is to also make the content of the `prov:Entity` nodes available, and ultimately the computational processes that underly the `prov:Activity` nodes. We hope that the clickable provenance graph we made available at `http://www.few.vu.nl/~michielh/lisc2013/prov/` serves as an example.

Acknowledgments

This work was funded under the Dutch COMMIT program as part of the Data2Semantics and SEALINCMedia projects.

References

1. H. Akil, M. E. Martone, and D. C. Van Essen. Challenges and opportunities in mining neuroscience data. *Science*, 331(6018):708–712, 2011.
2. A. Bate, M. Lindquist, I. Edwards, S. Olsson, R. Orre, A. Lansner, and R. De Freitas. A bayesian neural network method for adverse drug reaction signal generation. *European journal of clinical pharmacology*, 54(4):315–321, 1998.

3. K. Belhajjame, O. Corcho, D. Garijo, J. Zhao, P. Missier, D. Newman, R. Palma, S. Bechhofer, E. Garcia-Cuesta, J. Gmez-Prez, G. Klyne, K. Page, M. Roos, J. Ruiz, S. Soiland-Reyes, L. Verdes-Montenegro, D. De Roure, and C. Goble. A workflow-centric research objects: A first class citizen in the scholarly discourse. In *Proceedings of the ESWC2012 Workshop on the Future of Scholarly Communication in the Semantic Web (SePublica2012)*, Heraklion, Greece, May 2012.
4. E. Deelman, D. Gannon, M. Shields, and I. Taylor. Workflows and e-science: An overview of workflow system features and capabilities. *Future Generation Computer Systems*, 25(5):528–540, 2009.
5. S. J. W. Evans, P. C. Waller, and S. Davis. Use of proportional reporting ratios (prrs) for signal generation from spontaneous adverse drug reaction reports. *Pharmacoepidemiology and Drug Safety*, 10(6):483–486, 2001.
6. D. Garijo, P. Alper, K. Belhajjame, O. Corcho, Y. Gil, and C. Goble. Common motifs in scientific workflows: An empirical analysis. In *8th IEEE International Conference on eScience*, USA, 2012. IEEE Computer Society Press.
7. A. Goald. Practical pharmacovigilance analysis strategies. *Pharmacoepidemiology and drug safety*, 12(7):559–574, 2003.
8. C. Goble, R. Stevens, D. Hull, K. Wolstencroft, and R. Lopez. Data curation + process curation=data integration + science. *Briefings in Bioinformatics*, 9(6):506–517, 2008.
9. P. Groth and J. Frew. Proceedings of the 4th international conference on provenance and annotation of data and processes. 2012.
10. P. Groth, Y. Gil, and S. Magliacane. Automatic metadata annotation through reconstructing provenance. In *Third International Workshop on the role of Semantic Web in Provenance Management, ESWC 2012*, 2012.
11. P. Groth and L. Moreau. PROV-Overview: An Overview of the PROV Family of Documents. Working group note, W3C, Apr. 2013. `http://www.w3.org/TR/2013/NOTE-prov-overview-20130430/`. Latest version available at `http://www.w3.org/TR/prov-overview/`.
12. K. Kadoyama, I. Miki, T. Tamura, J. Brown, T. Sakaeda, and Y. Okuno. Adverse event profiles of 5-fluorouracil and capecitabine: Data mining of the public version of the fda adverse event reporting system, aers, and reproducibility of clinical observations. *International Journal of Medical Sciences*, 9(1):33–39, 2012.
13. M. Liu, M. E. Matheny, Y. Hu, and H. Xu. Data mining methodologies for pharmacovigilance. *ACM SIGKDD Explorations Newsletter*, 14(1):35–42, 2012.
14. J. P. Mesirov. Accessible reproducible research. *Science*, 327(5964):415–416, 2010.
15. H. A. Piwowar, R. S. Day, and D. B. Fridsma. Sharing detailed research data is associated with increased citation rate. *PloS one*, 2(3):e308, Jan. 2007.
16. K. Rothman, S. Lanes, and S. Sacks. The reporting odds ratio and its advantages over the proportional reporting ratio. *Pharmacoepidemiology and drug safety*, 13(8):519–523, 2004.
17. A. Szarfman, S. Machado, and R. O'Neill. Use of screening algorithms and computer systems to efficiently signal higher-than-expected combinations of drugs and events in the us fda's spontaneous reports database. *Drug safety : an international journal of medical toxicology and drug experience*, 25(6):381–392, 2002.
18. E. van Puijenbroek, A. Bate, H. Leufkens, M. Lindquist, R. Orre, and A. Egberts. A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. *Pharmacoepidemiology and drug safety*, 11(1):3–10, 2002.

Appendix A

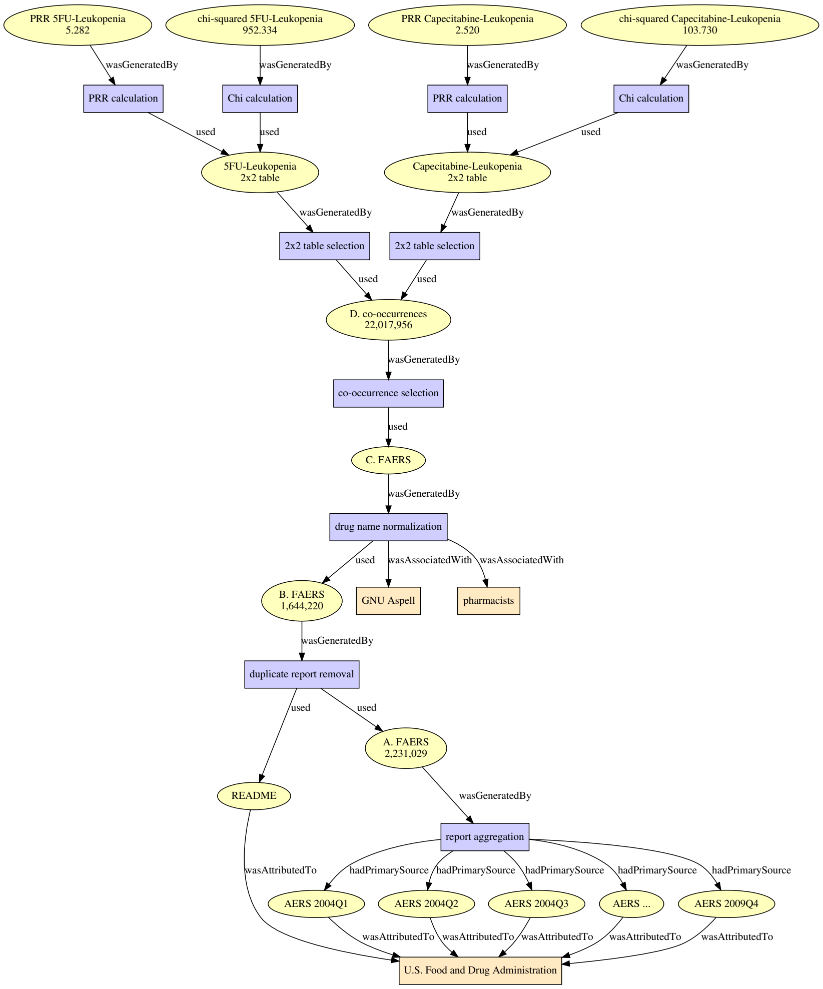

Fig. 1. Reproduction of the provenance graph corresponding to the computational workflow described in [12].