OpenVigil – Data quality and cleansing procedures

Ruwen Böhm <ruwen.boehm@pharmakologie.uni-kiel.de>
Hans-Joachim Klein <hjk@is.informatik.uni-kiel.de>
Version 2015-03-08

Introduction

OpenVigil 1 (http://www.uni-kiel.de/pharmacology/pvt/openvigil.php) is a pharmacovigilance data analysis tool. It is superseded by OpenVigil 2 (http://www.is.informatik.uni-kiel.de:8503/OpenVigil/) which is faster and more suited for data analysis since it operates on cleansed data. OpenVigil 1 is thus now deprecated for pharmacovigilance analyses but still maintained for exploring the raw data.

The data currently used in OpenVigil are taken from Adverse Event Reporting System (AERS) of the Food and Drug Administration (FDA) of the USA.

The advantage of the FDA source is a large amount of data due to the size of the reporting population. The disadvantage on the other hand is that reports in the AERS are often incomplete (e.g., missing patient demographic data) or wrong (e.g., non-professional reporter or biased reporting, see the OpenVigil cave-at documents).

Data quality of FDA AERS pharmacovigilance data

Missing or malformatted records

Raw FDA AERS ASCII quaterly data files contain various pitfalls:

- Some text lines which represent single records were accidentally broken down to two lines.
- Some text lines (= records) are cut off in the middle and are thus incomplete records, the next line belongs already to another record.

- Illegal characters at the beginning of a data file exist that might stop ASCII parsers from recognizing this file.
- Illegal characters that could break import into a SQL database exist.

Examples of import errors in **OpenVigil 1** are recorded in http://www.uni-kiel.de/pharmacology/pvt/openvigil.php?cd=if (fig. 1).

FNAME	DT	NERR_PARSER	NERR_SQL	TERR_PARSER	TERR_SQ
DEMO09Q3	2014-09-09 14:20:05	2	0	Parser error: item counts (expected 23/found 11) at line 53920, file ascii/DEMO09Q3.TXT do not match: 6307398$6784218$I$$6307398-8$20040501$20081001$20090810$PER$US-JNJFOC-20081000321$JOHNSON + JOHNSON PHARMACEUTICAL Parser error: item counts (expected 23/found 13) at line 53921, file ascii/DEMO09Q3.TXT do not match: 13YRFN$$$20090713$CN$$$$UNITED STATES	

Fig. 1: Example of errors found in the import log

OpenVigil 2 stores information about problems with the processing of raw data in internal tables which can only be accessed by using the SQL interface. During import it is possible to correct data using a comfortable interface (fig. 2, see below). Furthermore, drugname mapping and calculated values (age, daily dosage, therapy duration) can be inspected for all imported data.

"Drugname"

OpenVigil relies on the data field "drugname" which was conceived by the FDA to hold a text string that describes the medication used in this report. In the majority of cases, the supplied drugnames are easily understandable for humans and computer programs alike:

```
COUMADIN (WARFARIN SODIUM)
WARFARIN
WARFARIN POTASSIUM
```

There are, however, many inaccurate entries causing problems. Below are some examples of names that are either problematic for parsing because of the formatting or generally unusable due to ambiguity (e.g., conflicting information) or missing information:

```
BRODIFACOUM (SUPERWARFARIN )
COUMADIN (WARFARIN SODIUM) (5 MILLIGRAM) (WARFARIN SODIUM)
WARFARIN (WARFARIN /00014801/)
RIVAROXABAN 20MG OD OR WARFARIN OD (1, 2.5 OR 5MG)
BLOOD THINNER (NON-ABBOTT)
UNSPECIFIED ANTIVITAMIN K DRUG
480 10ML (LIPIODOL ULTRA FLUIDE) (ETHIODIZED OIL)
(RHO (D) IMMUNE GLOBULIN INTRAVENOUS (HUMAN)) LOT# 4344400001
(THIOPENTONE /00053401/)
(THERAPEUTTC RADIOPHARMACEUTICALS)
ADDERALL XR (AMFETAMINE ASPARTATE, AMFETAMINE SULFATE,
DEXAMFETAMINE
ACCU-CHEK CV TEST STRIP
ACETAMINOPHEN\TRAMADOL HYDROCHLORIDE
ALL OTHER THERAPEUTIC PRODUCTS
CC-5013 (LENALIDOMIDE ) (CAPSULES)
DECONGESTANTS AND ANTIALLERGICS(NO INGREDIENTS/SUBSTANCES)
'MULTIPLE' MEDICATIONS (ALL OTHER THERAPEUTIC PRODUCTS)
'NEW' ACE INHIBITOR
(ABH) ATIVAN, BENADRYL AND HALDOL
# 40 TYLENOL # 3
```

```
"breathing machine" when needed
```

Entries might also contain references to unknown or blinded study drugs, so even humans could not guess what was applied. There are many ambiguous reports like „WARFARIN BLINDED“ or „UNKNOWN“ that can never be resolved to a unique drugname or brandname.
The last example in the OpenVigil tutorials shows some common problems and pitfalls.

“Drugname” is different from the term “drug” which we use for a substance in a pharmaceutical product that is biologically active and responsible for the therapeutic effect. “Drug”, in turn, must not be confused with other meanings like illicit drugs or a ready-made pharmaceutical product like a pill, denoted by its brandname.

Examples of differences between USAN and other drug names

Because OpenVigil uses the U.S. American pharmacovigilance data, most drugs are named according to the U.S. Adopted Name (USAN) scheme. This differs from International Nonproprietary Name (INN):

International Nonproprietary Name (INN)	U.S. Adopted Name (USAN)
glibenclamide	glyburide
acetylsalicylic acid	aspirin
metamizole	dipyrone
salbutamol	albuterol
paracetamol	acetaminophen
rifampicin	rifampin
suxamethonium	succinylcholine
glyceryl trinitrate	nitroglycerin

Note that there are also other drugnames like the British Adopted Name (BAN) which exist in the raw FDA data. BAN allows combining two drugs into one “drugname”, e.g., cotrimoxazole as a combination of trimethoprim and sulfamethoxazole.

Dosages

Dosages can be reported in a variety of ways, e.g.,

```
10 MG BID ORAL
DURING THE THIRD TERM OF PREGNANCY
10DROP THREE TIMES PER DAY
10MG PER DAY
^FOR A COUPLE OF YEARS^
2.5-5MG AS NECESSARY
150 MG 1 X PER 1 DAY, ORAL
1MG IV Q4HOUR PRN; 1MG IV Q8HOUR PRN; 2 MG IV Q6HOUR PRN; 2MG
IV QHS^
```

Route of administration

While a very limited set of keywords is used here, some are redundant, e.g., “OCCLUSIVE DRESSING” and “OCCLUSIVE DRESSING TECHNIQUE” or “INTRAUTERINE” and “INTRA-UTERINE”.

Age and dates

Data quality of dates and patient ages is rather high. Still, single reports are probably wrong, e.g., “7200 YR” appears a bit old for a human while “109 YR” might be a valid report.

A simple logic to calculate various units (years, months, days) to an uniform format is required, e.g. “26983 DY” to years or vice versa.

Improving data quality

Fixing broken records

OpenVigil 1 does not offer any means to fix import errors. However, you are informed of the amount of data that could not be imported properly (fig. 1).

OpenVigil 2 offers manual correction of broken records (fig. 2a), entering new records (fig. 2b) and checking for duplicates (fig. 2c) .

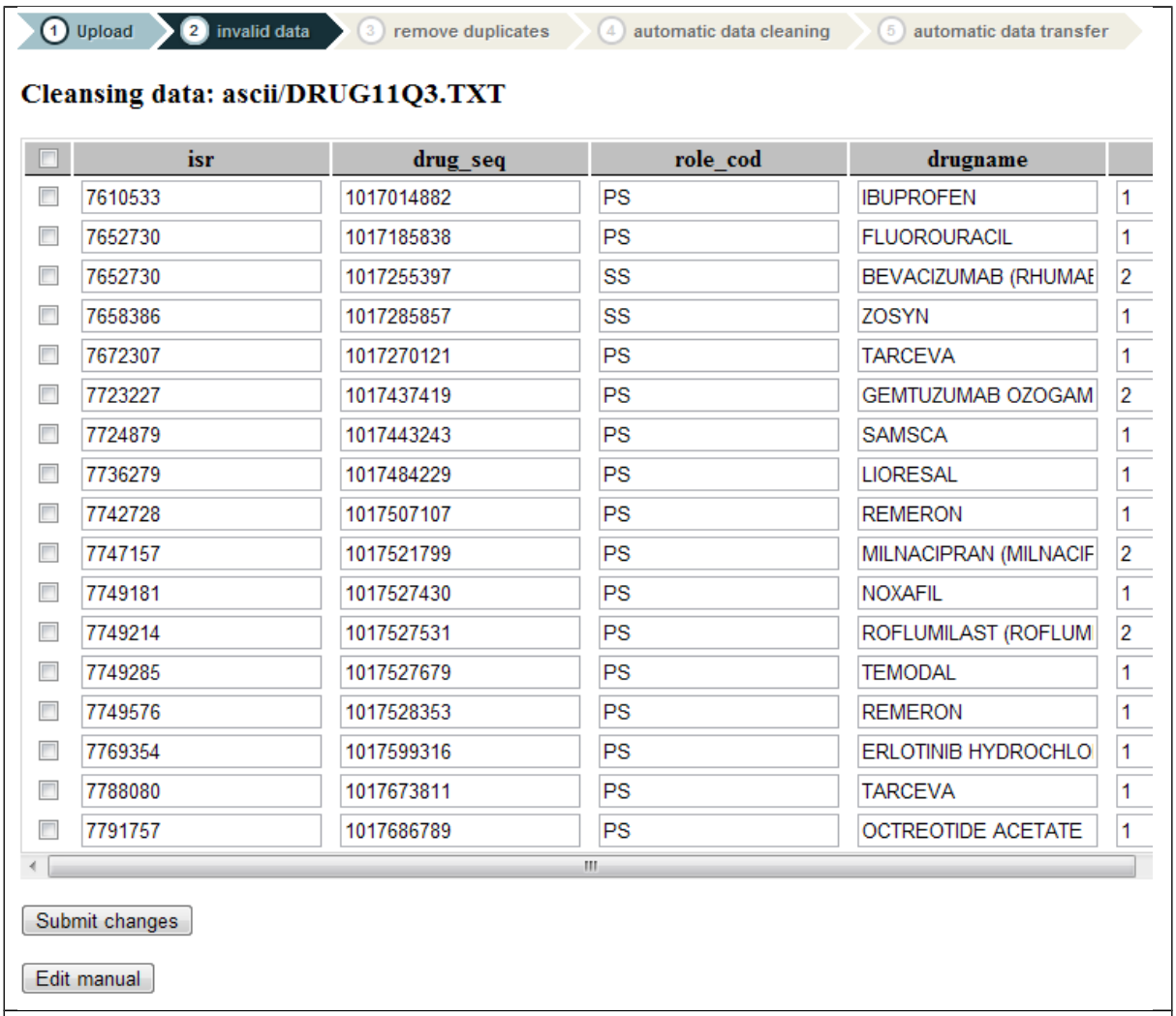

Fig. 2a: Manual data correction in OpenVigil 2 – broken ASCII lines
Several records of the ASCII raw data were split on two instead of one line. E.g., the first two errornous lines are:

```
7610533$1017014882$PS$IBUPROFEN$1$ORAL$78; 1X;PO
$$$$$074978$
```

OpenVigil automatically merges these lines and asks the importing user whether the merge is ok.

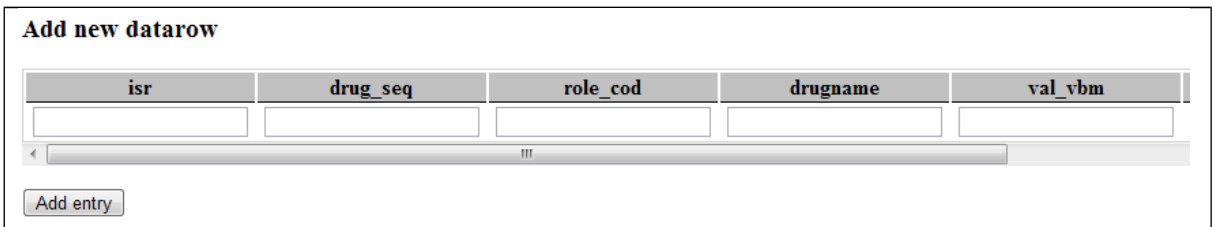

Fig. 2b Manual data correction in OpenVigil 2 – enter new records
It is also possible to enter new records if OpenVigil could not offer a good suggestion how to fix the broken record.

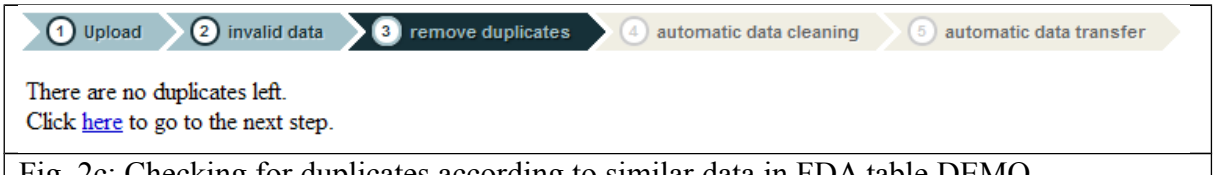

Fig. 2c: Checking for duplicates according to similar data in FDA table DEMO

(demographic data, e.g., patient age and sex).

Drugname mapping

The FDA AERS pharmacovigilance data contain the item DRUG.DRUGNAME. This verbatim, free-text textstring can most times be converted into a INN or USAN drugname using drug databases like Drugbank (http://www.drugbank.ca/downloads/archived), Drugs@FDA (http://www.fda.gov/Drugs/InformationOnDrugs/ucm079750.htm) or RXNORM (http://www.nlm.nih.gov/research/umls/rxnorm/docs/rxnormfiles.html or online at http://rxnav.nlm.nih.gov/RxNormRestAPI.html).

OpenVigil 1 prior to 1.2.7 does no drugname mapping but works with original, verbatim free-text FDA drugnames. **OpenVigil 1.2.7** introduces experimental drug-mapping via RXNORM. However, RXNORM will causes mismappings, e.g., „WARFARIN BLINDED“ is mapped to „WARFARIN“, so be very, very careful!

OpenVigil 2 does only unambiguous drugname-mapping (using Drugbank and a fallback to Drugs@FDA if the former does not suffice) of reports and is thus safe to use.
See the last example of the tutorials for the various pitfalls you can step into!

The mapping logic is presented in Eggeling 2013. The mapping process flow is roughly as follows:

- Entries in the drugname field of raw FDA data consisting of several parts like YASMIN (DROSPIRENONUM, ETHINYLESTRADIOLUM)' are decomposed into their components. Here is an example, how regular expressions are used to split the verbatim drugname text-string:

```
([ ]+(and|with|\+)[ ]+|[,/](?!([0-9]|ml|mg|m2|kg))| w/)
```

- Numbers with no obvious meaning like '/00599201/' are removed.
- For each component an assignment to a single drugname is looked for. If this is not possible an assignment to a brandname (pharmaceutical product) is tried. In both cases tables built using data from the Drugbank and Drugs@FDA are used in this order. If an exact match with the primary name is not possible, synonyms are checked for a possible match as well. A table with misspellings is consulted if an exact match fails.
- Components which could not be assigned to drugnames or brandnames in the preceding step are analysed for denoting a drug in salt form. It is tried to reduce it to a basic name known as drugname or synonym of a drugname, e.g., 'METFORMIN HYDROCHLORIDE to 'METFORMIN'. If this can be done with the help of a table containing typical salt designators, the component is mapped to the combination of basic name and salt designator and stored as a synonym of the drugname.
The salt table (from A like acetate to V like versenate) can be found in the source or the WAR file at *SQL/Salts.csv*.

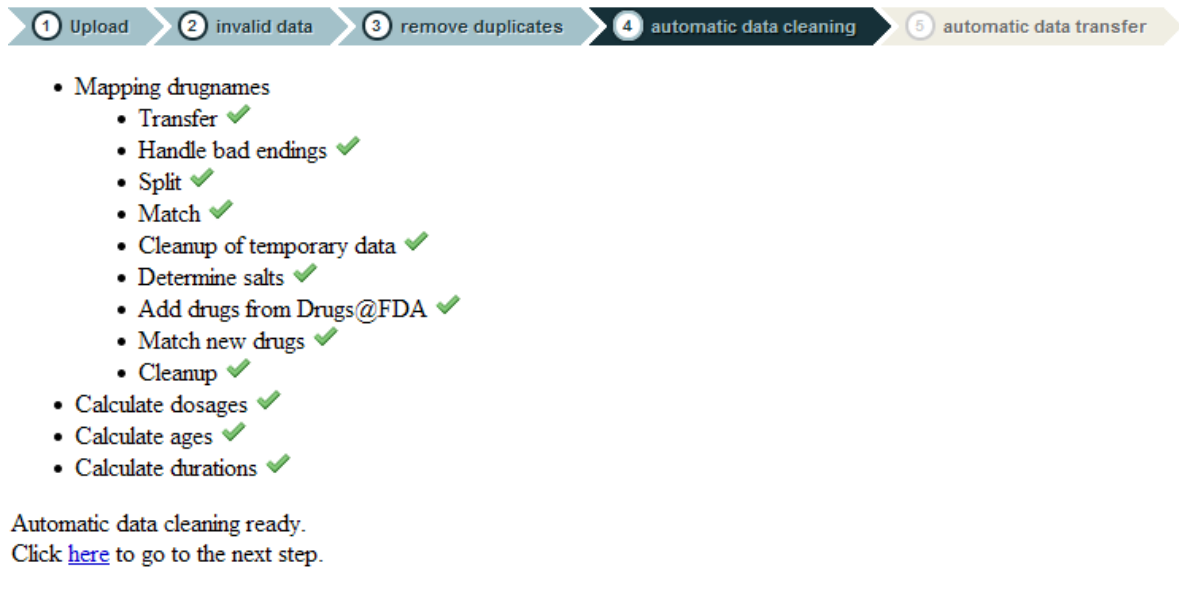

Fig. 2d: Mapping verbatim, free-text FDA DRUG.DRUGNAME to unique, unambiguous INN/USAN drugnames using Drugbank oder Drugs@FDA data.

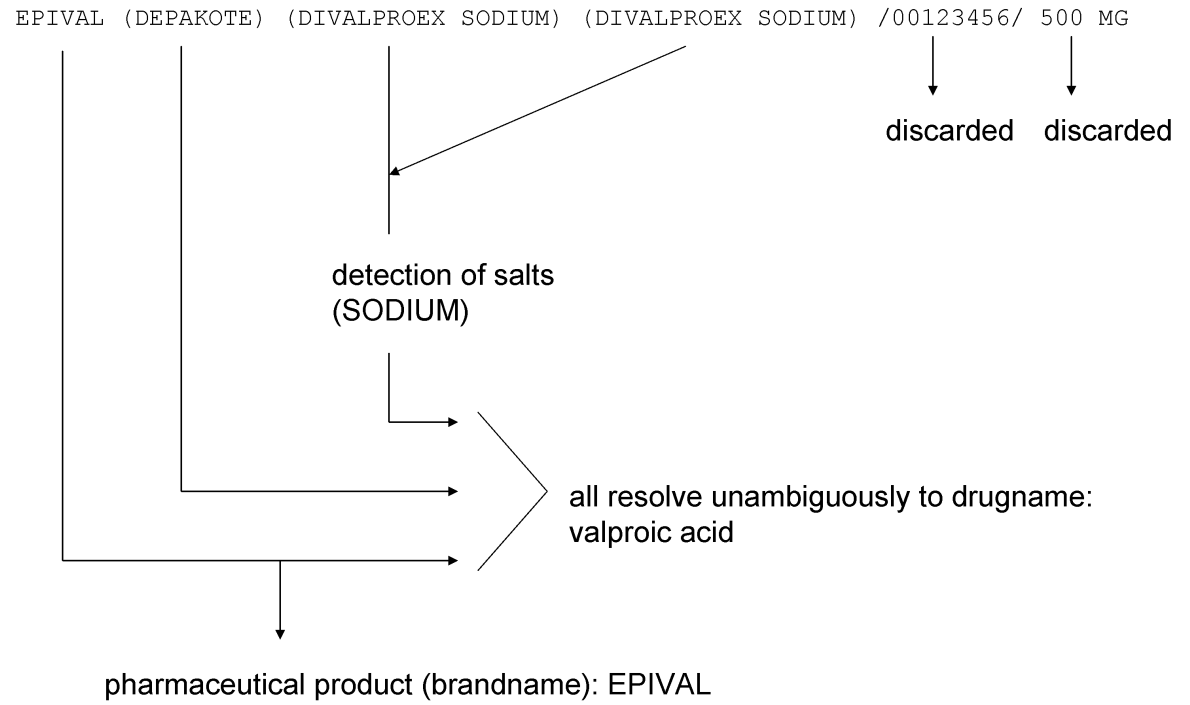

Fig. 3: Using drugname-mapping data, a list of salts and one regular expression, even complicated verbatim, free-text "drugnames" are mapped unambiguously to a USAN/INN drugname (and, in this case, additionally to a pharmaceutical product).

Calculating dosages and ages

OpenVigil 1 does currently not provide any calculation logic. OpenVigil 2 can parse and calculate dosages and ages:

For the calculation of daily dosage in mg, the regular expressions in table 1 are used.

Tab. 1: Regular expressions for dosage calculations (Eggeling 2013)	
once daily	`^[0-9]+([.,]{1}[0-9]+)?[]* (MG|MILIGRAM|MILIGRAMS|MILLIGRAM|MILLIGRAMS) [ ,;]* ((1 DAY)|(ONE DAILY MORNING)|(ONCE A DAY)|(QD)|(ONCE DAILY)|(DAILY( []*[[.(.]]?(1/D)[[.).]]?)?)|(PER DAY)| (1X/DAY))+$`
twice daily	`^[0-9]+([.,]{1}[0-9]+)?[]* (MG|MILIGRAM|MILIGRAMS|MILLIGRAM|MILLIGRAMS)[ ,;]* (((TWICE)(PER|A|(IN A)) DAY)|((TWICE) DAILY)|(DAILY[ ]*[[.(.]]?(2/D)[[.).]]?)|(2X/DAY))+$`
three times daily	`^[0-9]+([.,]{1}[0-9]+)?[]* (MG|MILIGRAM|MILIGRAMS|MILLIGRAM|MILLIGRAMS) [ ,;]*(((THRICE)(PER|A|(IN A)) DAY)|((THRICE) DAILY)| (DAILY[ ]*[[.(.]]?(3/D)[[.).]]?)|(3X/DAY))+$`
four times daily	`^[0-9]+([.,]{1}[0-9]+)?[]* (MG|MILIGRAM|MILIGRAMS|MILLIGRAM|MILLIGRAMS) [ ,;]*((((FOUR TIMES)) (PER|A|(IN A)) DAY)|(((FOUR TIMES)) DAILY)|(DAILY[ ]*[[.(.]]?(4/D)[[.).]]?)|(4X/DAY))+$`

not calculable	`^(((DAILY)|(TEXT)|(DOSE))[: ]*)*((UNK)|(UKN)| (UNKNOWN)|(UNKNOWN DOSE)|(UNSPECIFIED)|(DOSING INFORMATION UNKNOWN)|(DOSAGE IS UNCERTAIN)|(AS NEEDED)|(AS REQUIRED)|(NOT REPORTED)|(NOT PROVIDED)| (DF OTHER)|,|\\.| |\\(|\\)|;)+$`
	
	`^[ .,-]*$`
	
	`^[0-9]+([.,]1[0-9]+)?[ ]*(MG|MILIGRAM|MILIGRAMS| MILLIGRAM|MILLIGRAMS)$`

References

Eggeling Ch. [Data quality in pharmacovigilance data] Datenqualität in Pharmakovigilanzdaten. Master Thesis 2013 http://www.is.informatik.uni-kiel.de/~hjk/masterarbeit_Eggeling.pdf