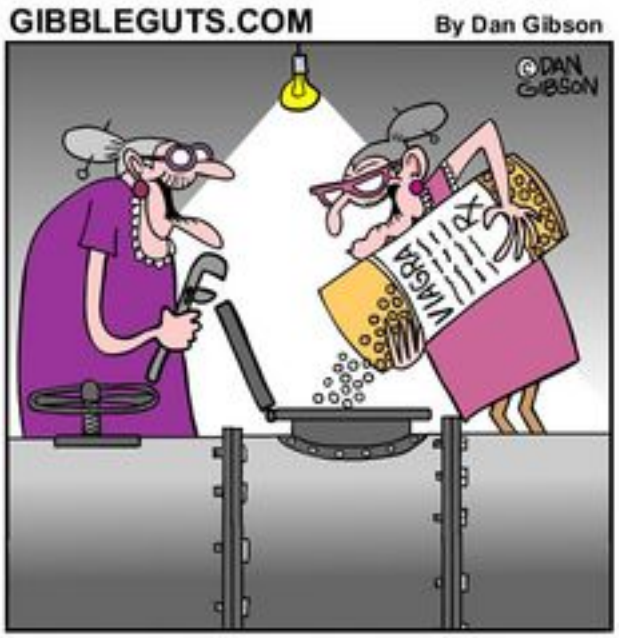
Study about the legal use of sildenafil and methylphenidate

Report of the work experience internship

Academic year 2014-2015



Life was boring around the old folks home until Gert and Tillie started spiking the water supply with Viagra.

Name: Irene Zanellato

Student number: s1406248

Internal supervisor: Prof. Marta Fiocco

External supervisor: Dr. Bastiaan J. Venhuis

Contents

1. RIVM (Rijksinstituut voor Volksgezondheid en Milieu)

The internship experience took place at the RIVM, the National Institute of Public Health and Environment situated in Bilthoven. It represents the main public institute which aims to continuously detect the public health in lots of different spheres of interest. It works for the government authorities in order to develop suitable policy, mainly in the field of public health, nutrition, safety and environmental management. It has been designated as a WHO (World Health Organization) collaborating center and it also collaborates with other several international organizations. The work that is done at RIVM is also recognized all over the world, especially for those researches in fields concerning nutrition, tobacco product regulation and immunotoxicology.

The main RIVM's clients are the Ministry of Health, Welfare and Sport (VWS), the Ministry of Infrastructure and the Environment (I&M), the Ministry of Economic Affairs, Agriculture and Innovation (EL&I) and the Ministry of Social Affairs and Employment (SCW), in addition to some international organizations like the European Union, the World Health Organization (as mentioned earlier) and the United Nations.

There are four major technical divisions at RIVM, which are in turn divided into several subsidiary centers or laboratories:

1. infectious diseases control
    - center of infectious diseases control
    - national coordination center for outbreak management
    - diagnostic laboratory for infectious diseases and perinatal screening
    - laboratory for vaccine-preventable diseases
    - microbiology laboratory for health protection
2. environment and safety
    - advisory service for the inspectorate
    - center for external safety
    - laboratory for ecological risk assessment
    - laboratory for radiation research
    - laboratory for environmental monitoring
    - center for environmental health
    - expertise center for substances
    - national poisons information center
3. public health and health services
    - youth health center
    - center for national screening programs
    - expertise center for methodology and information services
    - center for public health forecasting
    - center for prevention and health services research
4. nutrition, medicines and consumer safety division
    - laboratory for food and residue analysis
    - center for biological medicines and medical technology
    - center for nutrition and health
    - center for quality and pharmaceutical products
    - center for substances and integrated risk assessment
    - laboratory for toxicology, pathology and genetics

I worked on a project at the pharmaceutical department (in the fourth division previously mentioned). I alternated work directly at RIVM site (especially when I had to use the local computer) and at home. When

I was working at the RIVM location, I was sharing a room with other 3-4 interns. I cannot say if I was lucky or if it is a general characteristic, but I found the environment extremely friendly, relaxed and social. Apart from initial difficulties with the communication (sometimes it is hard to breach the Dutch language barrier, prevalently during the big collective moments such as lunch), I perceived a very comfortable, chummy and warm atmosphere (maybe less warm in terms of real temperature).

2. Description of the project and activities performed

2.1 Introduction and goal

The data analyzed regard the legal usage of two different medicines: sildenafil and methylphenidate. The data have been collected through an automated system that registered every single purchase occurred in pharmacies sited in Amsterdam, Utrecht and Eindhoven. Such a system permits to establish the identity of the purchaser (identified with a unique code), the city in which he/she bought the medicine, the date in which the purchase has happened, the specific subtype of medicine bought (further details will be given later on) and the corresponding DDD value (details will also follow). The datasets used for the analysis cover a period running from 2012 to 2014 for sildenafil and from 2012 to 2013 for methylphenidate. This analysis runs in parallel to another for which information about the usage of these two medicines is gathered by searching specific chemical values to be found in the sewage. Information regards not only the legal usage, but also the illegal one, which allows to estimate by subtraction the percentage of illegal quantities purchased and used by people. Clearly even more interesting is to detect how much illegal sildenafil circulates.

Sildenafil is the most worldwide used medicine to cure the erectile dysfunction. Nevertheless, it was originated for a different purpose. Sildenafil was conceived in 1996 by a group of chemists working at the English headquarter of the pharmaceutical company Pfizer with the goal to cure the pulmonary arterial hypertension. Nowadays such a medicine is popularly named Viagra and it is sold in the pharmacies in the form of oral suspension or tablet of different quantities: tablet of 20mg or the oral suspension of 10mg/ml (also called revatio, a specific brand of sildenafil) is used for the pulmonary disease, while tablets of 25mg, 50mg or 100mg are used for the erectile dysfunction.

Methylphenidate is used to treat the attention deficit hyperactivity disorder (ADHD), postural orthostatic tachycardia syndrome and narcolepsy. The Novartis Corporation owns the patent. The use of methylphenidate started in 1955 in order to treat the hyperactivity. For this medicine there are different types of administration; the following lists those taken into account in the analysis: capsule mga of 5mg, 10mg, 20mg, 30mg, 40mg, tablet of 5mg, 10mg, 20mg and tablet mga of 18mg, 27mg, 36mg, 54mg.

The main goal of the project is to derive an estimate for the quantity of legal usage, in order to later determine the quantity of illegal usage. Other relevant plots and estimates are of a certain interest. An in-depth analysis has already been done by some predecessors. The reason why I came in was to check whether there was any particular explanation for an interestingly anomalous result coming from the statistical analysis; but let's first introduce the concept of DDD.

The acronym DDD stays for "daily defined dose" and it is a statistical measure of drug consumption defined by the WHO; it indicates the *average maintenance dose per day for a drug used for its main indication in adults*. The DDD measurement doesn't strictly reflect the prescribed daily dose. Therefore DDD can be thought as representative of the global usage independently of the differences between patients and hence it is a rough estimate of consumption. Its calculation doesn't depend neither on the medicine prize, nor on the dosage form. One of the most interesting value produced by the analysis is the average daily use per person expressed in DDD (always referring to both medicines). Some calculations are necessary in order to derive such a quantity. Then a statistic of interest that can give us a rough idea of the overall daily use for the all population, not only distinguished by medicine, but also by dosage form must be defined. Such overall statistic has been calculated for the following subgroups:

- Sildenafil 10mg/ml, 20mg (revatio)
- Sildenafil 25mg, 50mg, 100mg
- Methylphenidate 5mg, 10mg, 20mg, 30mg, 40mg
- Methylphenidate 18mg, 27mg, 36mg, 54mg

So far, the mean has been chosen as summary statistic for all the subgroups. After a careful look, it appeared that the overall average daily use of revatio was 1.6 DDD, which is far greater than the expected 1.2 DDD generally prescribed by doctors. So, a main question arises: why do we have such a high value? What is a possible explanation for it?

2.2 Data manipulation and plot with the statistic of interest

The initial dataset has the following form:

patient	aflever_datum	stad	prk_naam	ddd
0006b0be534ea32c3cd1147512433288	2-10-2012	ams	SILDENAFIL TABLET 50MG	8
0006b0be534ea32c3cd1147512433288	21-2-2013	ams	SILDENAFIL TABLET 50MG	8
0006b0be534ea32c3cd1147512433288	20-6-2013	ams	SILDENAFIL TABLET 50MG	8
0006b0be534ea32c3cd1147512433288	15-4-2014	ams	SILDENAFIL TABLET 50MG	8
0007055afe14a6e07a99d71bdcabc76c	10-4-2013	ams	SILDENAFIL TABLET 100MG	8
0007055afe14a6e07a99d71bdcabc76c	15-10-2013	ams	SILDENAFIL TABLET 100MG	8

Figure 1: Initial data set

As previously mentioned, each patient is identified by a unique code (relabeled "ID"). Each patient can then appear more than once depending on the number of purchases he/she has done (relabeled "Date"). The cities are reported with abbreviations (relabeled "City"): ams for Amsterdam, utr for Utrecht and ein for Eindhoven. The prk_naam variable describes the type of medicine and the dosage form (relabeled "Dosage"). The last column represents the DDD value (labeled "ddd"). The main goal now is to derive the average daily use per subject. In order to do so we first need to determine a few additional quantities:

1. time interval from one purchase to the next one within each patient ("Periods"). This variable is calculated counting the number of days from the first purchase date to the second one, and then from the second to the third and so on. The numerical values obtained are then registered in order from the first patient's row in the data set to the row corresponding to the last line of the same patient. For the last purchase date there is no following date, so in correspondence to the last Periods there is always an unknown date (NA);
2. the numerical dosage value in correspondence to the type of medicine and dosage form ("Dos.in.mg"). This is simply derived from the categorical variable Dosage. For example the category *sildenafil tablet 20mg* has a corresponding value of 20 in Dos.in.mg. The only exception regards the category *sildenafil susp oraal 10mg/ml* which for medical reasons takes value 20 in Dosage.in.mg and not 10. Usually the doctor prescribes each time two oral suspensions of 10mg each;
3. number of tablets (or capsules, depending on the type of dosage) purchased by a subject ("num.of.tablets"). It is calculated by multiplying the ddd personal value with the ddd established value for the specific type of medicine and then by dividing for the dosage in mg. In formula:

- for sildenafil:
$$number\ of\ tablets = \frac{ddd * 50}{dosage\ in\ mg}$$
- for methylphenidate:
$$number\ of\ tablets = \frac{ddd * 30}{dosage\ in\ mg}$$

At this stage it is possible to calculate the average daily use per person both expressed in mg ("avg.daily.use.in.mg") and in ddd ("avg.daily.use.in.ddd"). Within each patient the average daily use is calculated per Periods, which means that there can be more than one average daily use value per subject. In formula:

- for sildenafil:
$$average\ daily\ use\ in\ mg = \frac{number\ of\ tablets * dosage\ in\ mg}{periods\ of\ time}$$

$$average\ daily\ use\ in\ ddd = \frac{average\ daily\ use\ in\ mg}{50}$$

- for methylphenidate:

$$average\ daily\ use\ in\ mg = \frac{number\ of\ tablets * dosage\ in\ mg}{periods\ of\ time}$$

$$average\ daily\ use\ in\ ddd = \frac{average\ daily\ use\ in\ mg}{30}$$

At this stage we have the following variables:

- `ID`: categorical
- `Date`: numeric of class 'Date'
- `City`: categorical
- `Dosage`: categorical
- `ddd`: numeric
- `Periods`: numeric
- `Dos.in.mg`: numeric
- `num.of.tablets`: numeric
- `avg.daily.use.in.mg`: numeric
- `avg.daily.use.in.ddd`: numeric

All these calculations are performed to estimate a statistic concerning the average daily use expressed in ddd. The choice made by my predecessor was to consider the mean of the avg.daily.use.in.ddd values. In order to discard any mistake that might have occurred during the computations, I started from the beginning by performing all required data preparation by myself instead of using the code of the previous student. This means that, starting from the initial dataset, I implemented an R code which creates the variables of interest previously described and plots the mean values in a graph. In this way I could compare the results obtained with procedure implemented as above described with the results based on the analysis performed by my predecessor. The two procedures gave exactly the same results, which means there were no mistakes in the calculation (at least supposing that the chance that two people writing two different R codes get the same mistake is close to zero); the problem concerning the presence of the odd result was still unsolved. Figure 2 illustrates the means with corresponding confidence intervals.

At this stage I decided to take one step back and check the distribution of the data for each subset of dosage type. All distributions are highly skewed, as can be noticed from Figures 3-6.

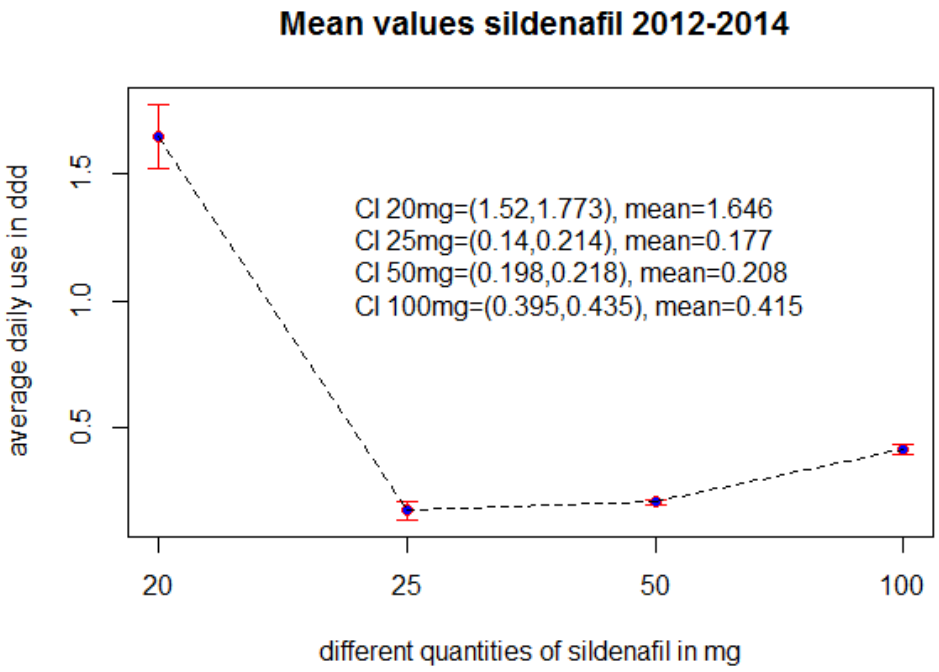


Figure 2: Means and corresponding CI for sildenafil

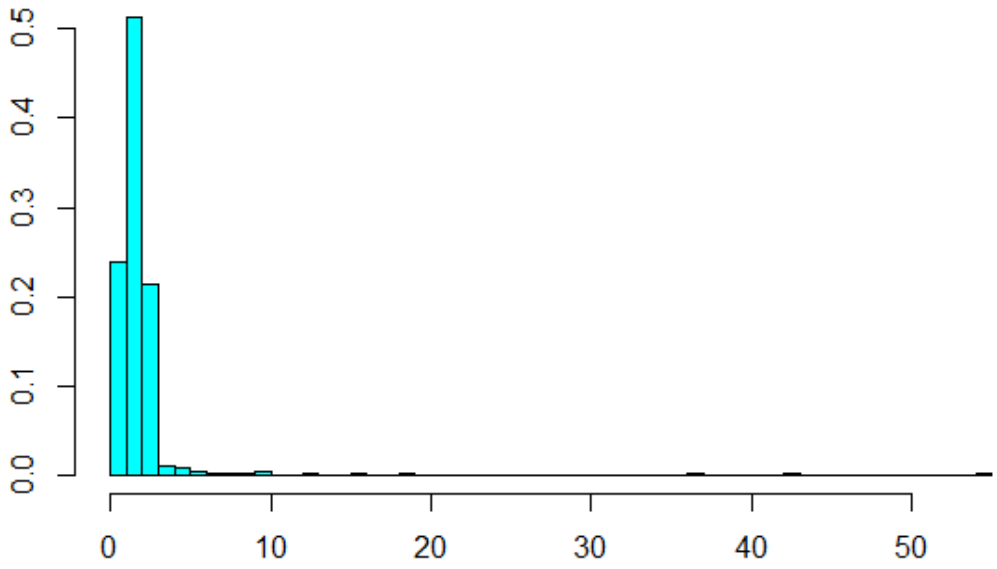


Figure 3: Histogram sildenafil 20 mg

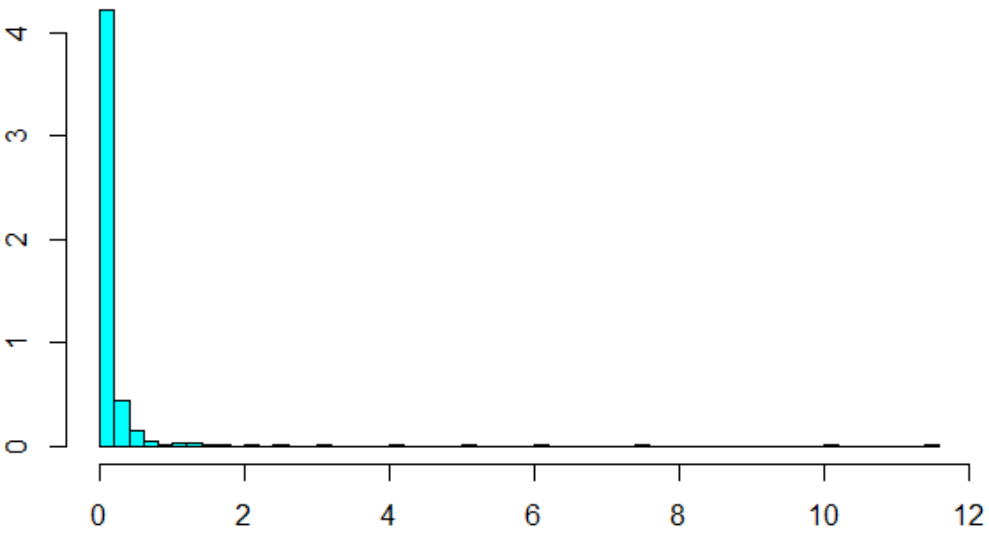


Figure 4: Histogram sildenafil 25 mg

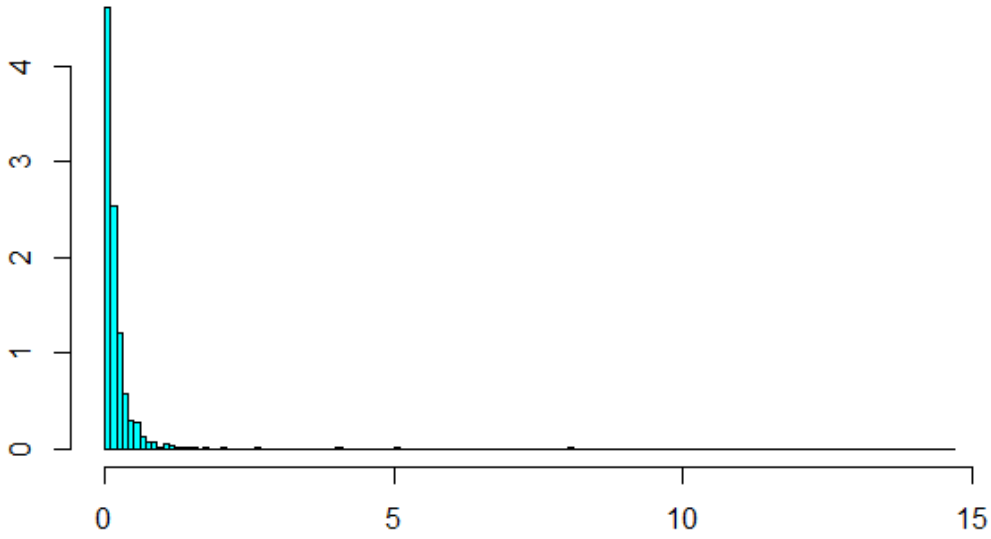


Figure 5: Histogram sildenafil 50 mg

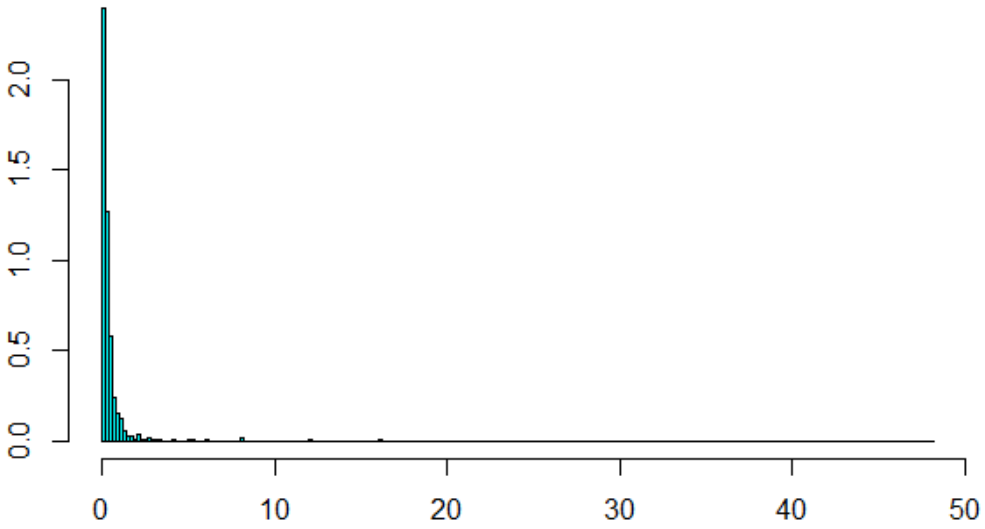


Figure 6: Histogram sildenafil 100 mg

- skewness sildenafil 20 mg: 13.585
- skewness sildenafil 25 mg: 11.308
- skewness sildenafil 50 mg: 13.792
- skewness sildenafil 100 mg: 17.998

The skewness is a measure of asymmetry from the normal distribution. Remember that a symmetrical distribution has a skewness of 0, an asymmetric distribution with a long tail to the right has a positive skewness and an asymmetric distribution with a long tail to the left has a negative skewness. According to Bulmer (1979) the rule of thumb for the skewness states that:

- if skewness is less than -1 or greater than +1, the distribution is highly skewed;
- if skewness is between -1 and $-\frac{1}{2}$ or between $+\frac{1}{2}$ and +1, the distribution is moderately skewed;
- if skewness is between $-\frac{1}{2}$ and $+\frac{1}{2}$, the distribution is approximately symmetric.

Now a question arises: in a strongly skewed distribution what is the best indicator of central tendency? According to Pagano and Gauvreau (2000) *"the best measure of central tendency for a given set of data often depends on the way in which the values are distributed. (. . .) When data are not symmetric, the median is often the best measure of central tendency. Because the mean is sensitive to extreme observations, it is pulled in the direction of the outlying data values, and as a result might end up excessively inflated or excessively deflated"*. So basing on the theoretical assumptions and what we observe in practice, the best choice for the indicator of central tendency is the median. Again the plot for different dosage forms, but this time considering the median instead of the mean.

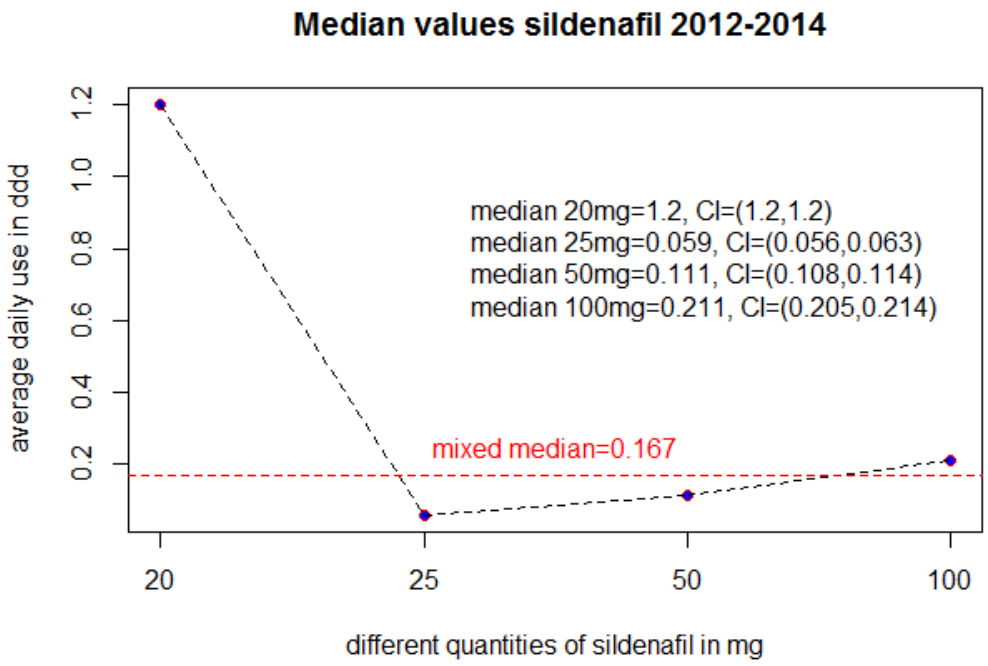


Figure 7: Medians and corresponding CI for sildenafil

As can be seen from Fig. 7, the median value for sildenafil 20 mg is exactly 1.2 as expected. The red dotted line represents the mixed median, which is the median calculated considering all the values together, independetly of the dosage type. The confidence intervals for the median have a different formulation compared to those for the mean. The confidence limits for a single median are given by actual values in the sample, which are chosen basing on the position given by the following rule based on a nonparametric procedure (a nonparametric method usually works efficiently when the normality assumption is violated):

- lower 95% confidence interval: $\left(\frac{n}{2} - \frac{1.96*\sqrt{n}}{2}\right) ranked\ value$
- upper 95% confidence interval: $\left(1 + \frac{n}{2} + \frac{1.96*\sqrt{n}}{2}\right) ranked\ value$

The function `ci.median` in `R` from library `asbio` is an alternative way to the hand calculation to compute the confidence interval for the median. In addition it allows to check which is the actual coverage of the interval, that will not be exactly 95% because the margins are numbers directly taken from the sample. In our case the actual coverage is:

- 95.16% for sildenafil 20 mg
- 95.60% for sildenafil 25 mg
- 95.19% for sildenafil 50 mg
- 95.06% for sildenafil 100 mg

The same procedure is then applied to other data regarding the use of methylphenidate. I first checked whether the distribution of the average daily use in ddd values had a skewed distribution or not. Once the former hypothesis was confirmed, I considered again the median as indicator of central tendency and I created the same type of plot. Concerning the methylphenidate data, I made some changes compared to what has been done before: initially the data have been gathered in groups defined by the dosage quantity in mg. However, from a medical point of view a capsule mga of 5 mg has a different purpose than a tablet of 5 mg (and the same way of thinking for the other quantities); consequently this time I grouped the average daily use values based on the dosage type, obtaining three different graphs: one for capsule mga, one for tablet and one for tablet mga, which follow together with the skewness values and the histograms.

- skewness methylphenidate capsule mga 5 mg: 7.304
- skewness methylphenidate capsule mga 10 mg: 14.99
- skewness methylphenidate capsule mga 20 mg: 10.136
- skewness methylphenidate capsule mga 30 mg: 12.586
- skewness methylphenidate capsule mga 40 mg: 8.016
- skewness methylphenidate tablet 5 mg: 12.713
- skewness methylphenidate tablet 10 mg: 19.494
- skewness methylphenidate tablet 20 mg: 1.671
- skewness methylphenidate tablet mga 18 mg: 19.4
- skewness methylphenidate tablet mga 27 mg: 18.203
- skewness methylphenidate tablet mga 36 mg: 17.929
- skewness methylphenidate tablet mga 54 mg: 14.483

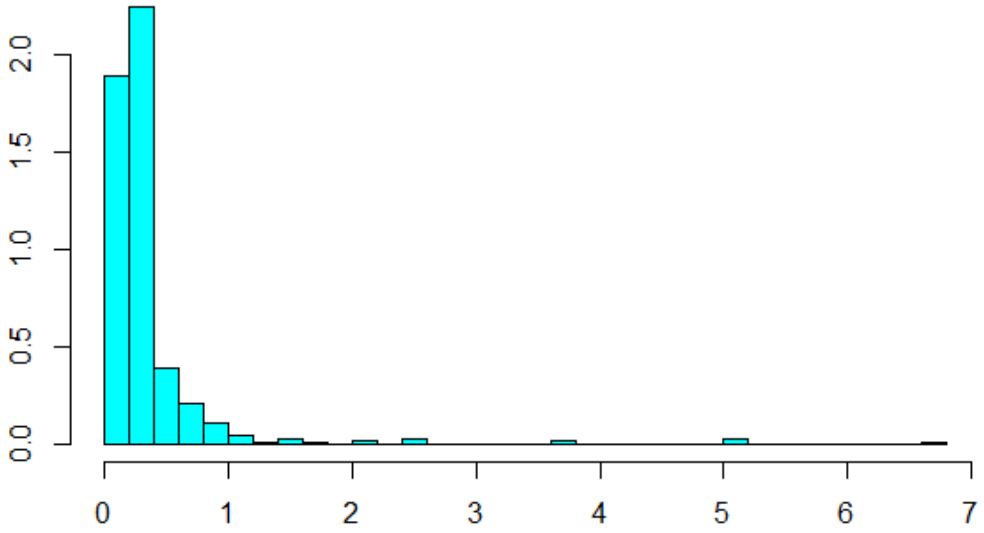


Figure 8: Histogram Methylphenidate capsule 5 mg

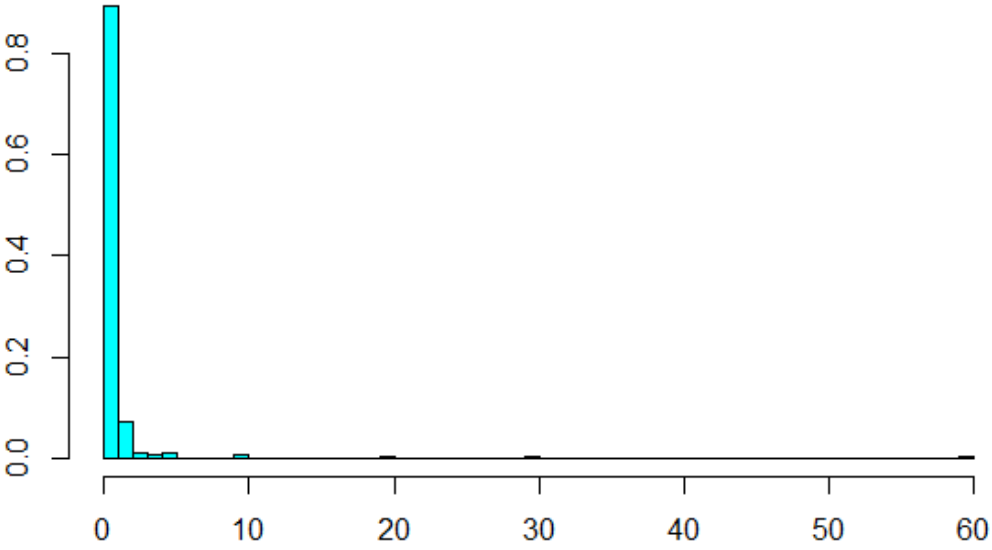


Figure 9: Histogram Methylphenidate capsule 10 mg

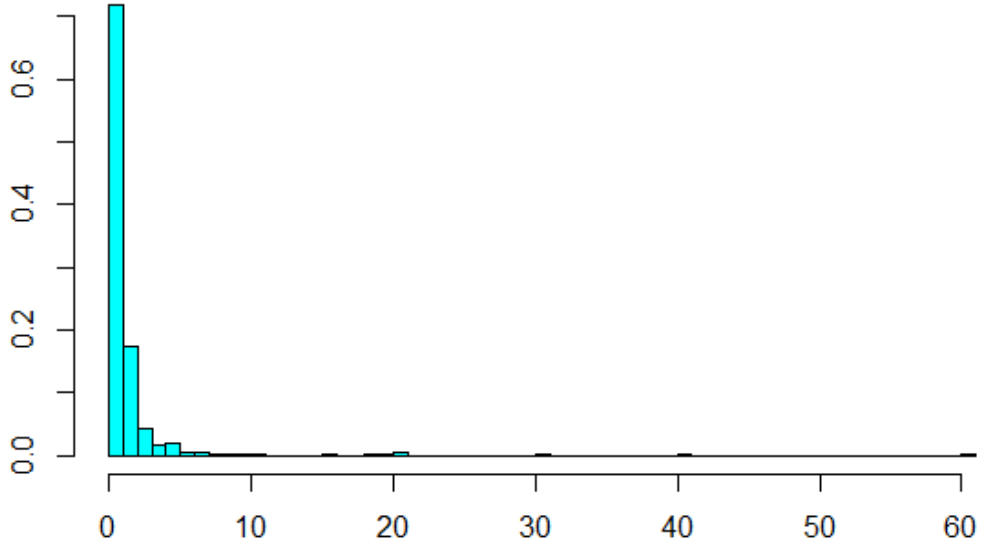


Figure 10: Histogram Methylphenidate capsule 20 mg

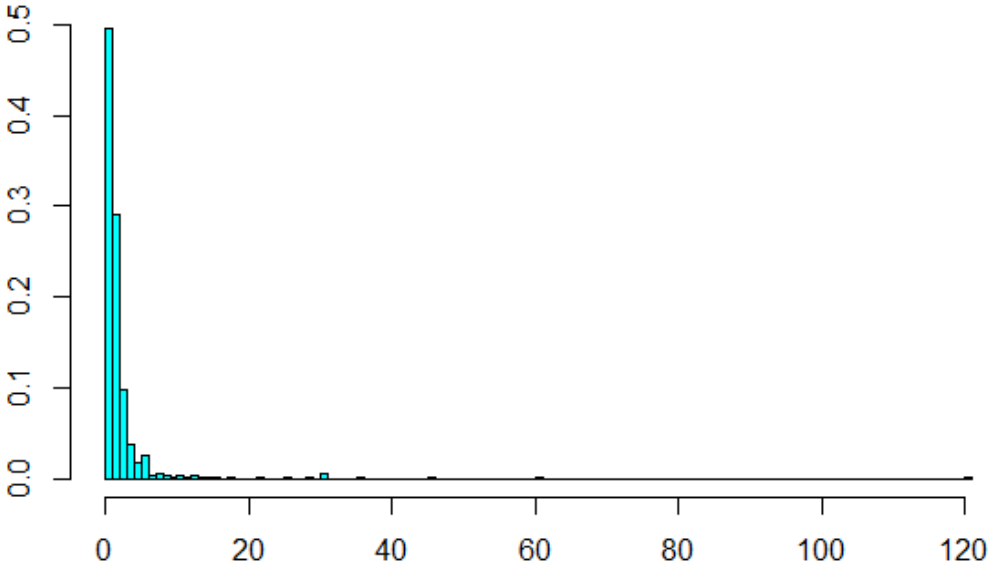


Figure 11: Histogram Methylphenidate capsule 30 mg

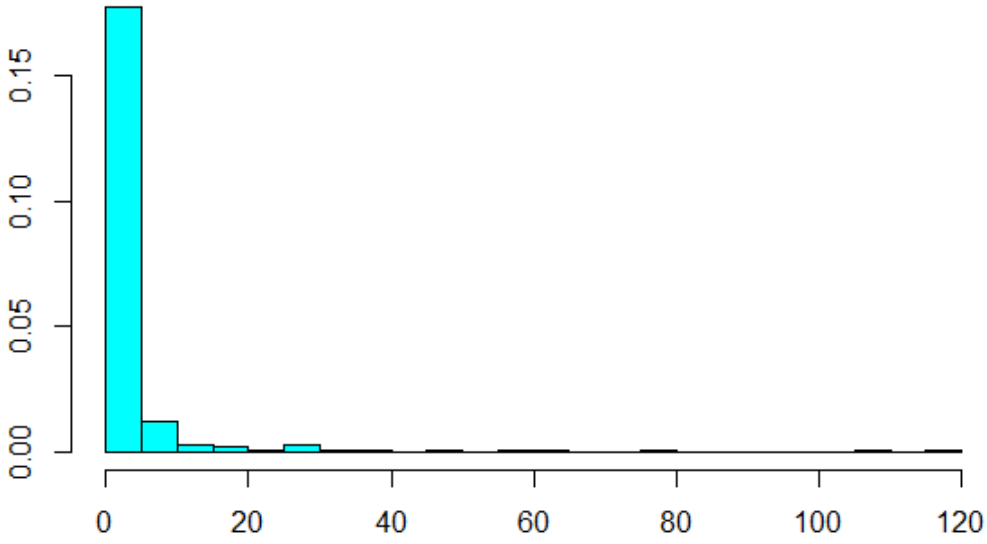


Figure 12: Histogram Methylphenidate capsule 40 mg

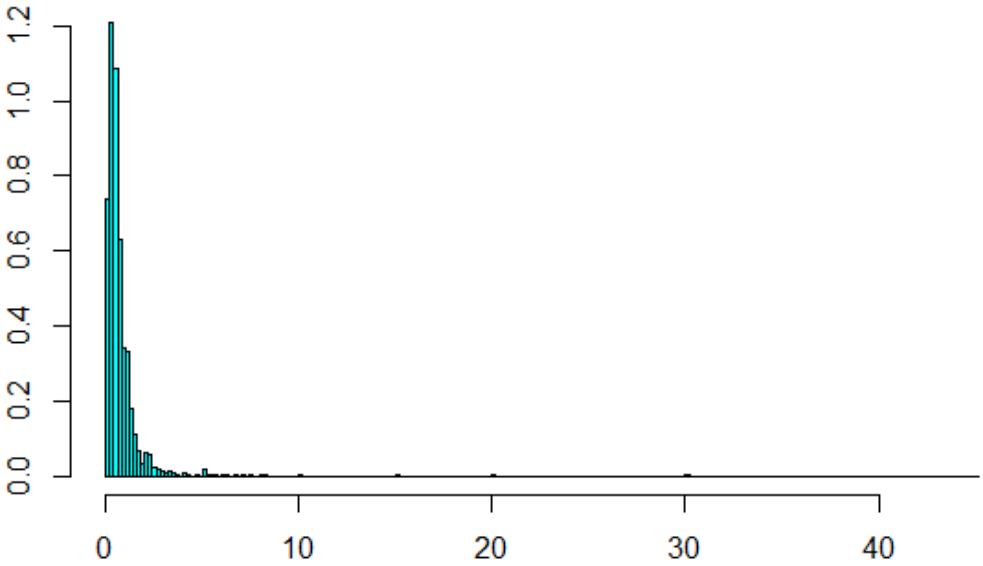


Figure 13: Histogram Methylphenidate tablet 5 mg

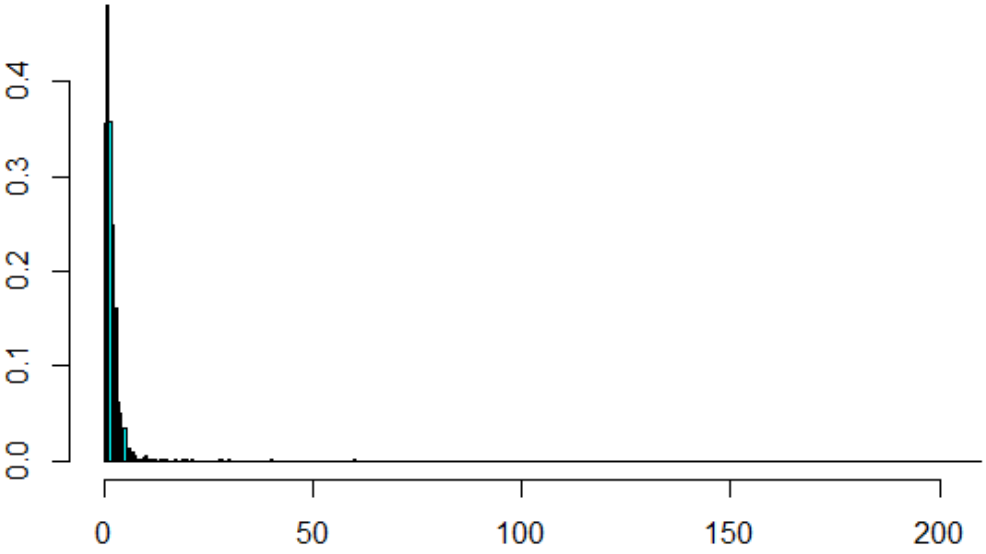


Figure 14: Histogram Methylphenidate tablet 10 mg

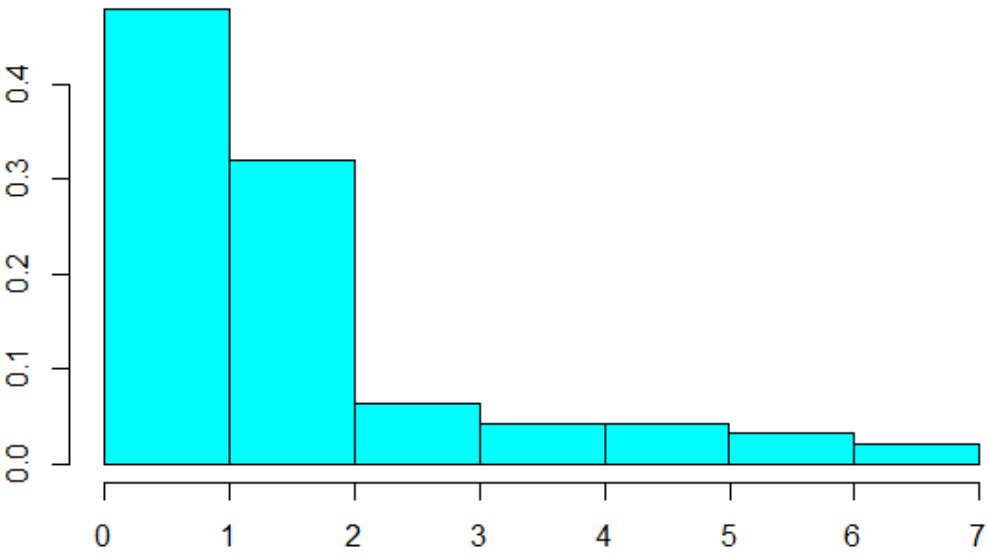


Figure 15: Histogram Methylphenidate tablet 20 mg

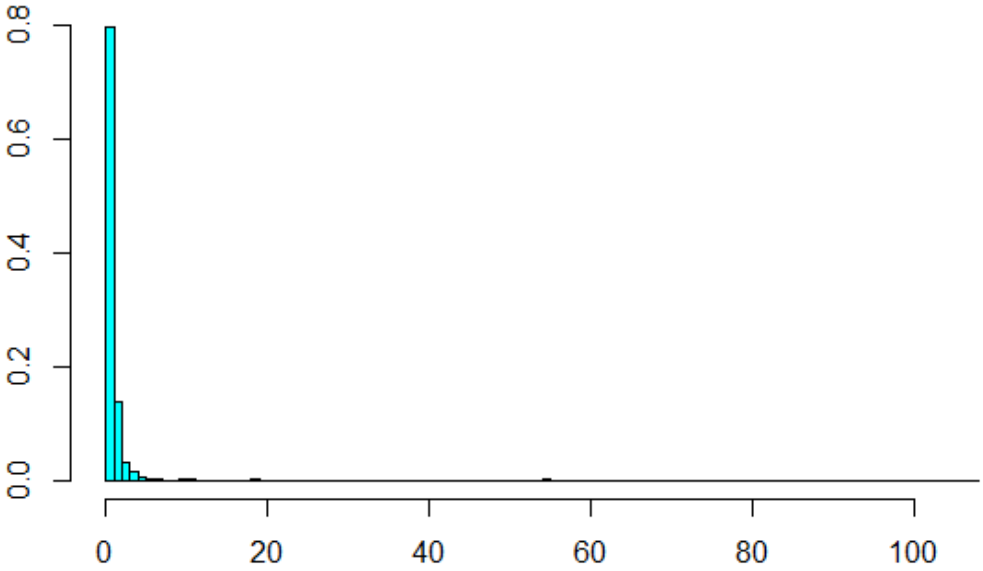


Figure 16: Histogram Methylphenidate tablet mga 18 mg

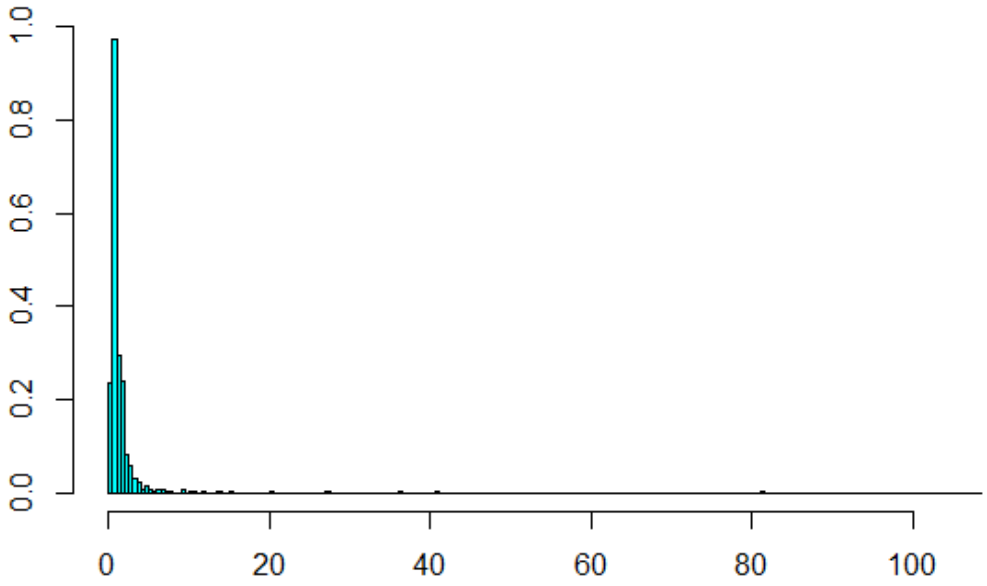


Figure 17: Histogram Methylphenidate tablet mga 27 mg

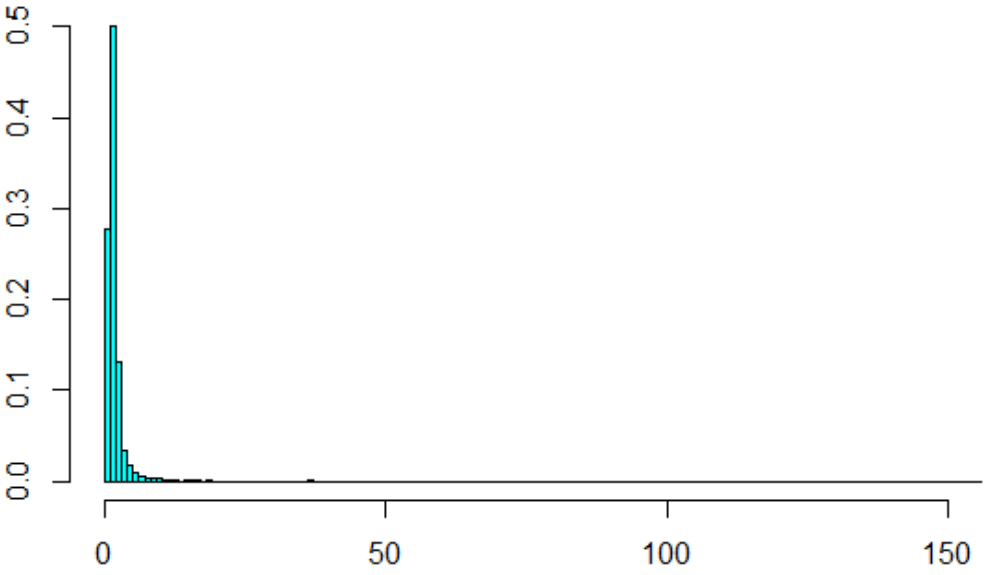


Figure 18: Histogram Methylphenidate tablet mga 36 mg

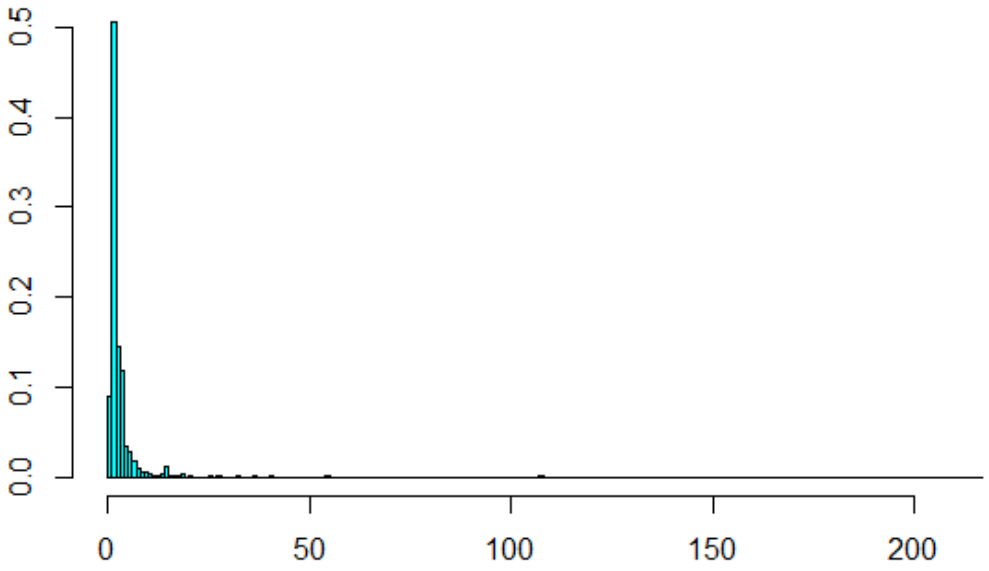


Figure 19: Histogram Methylphenidate tablet mga 54 mg

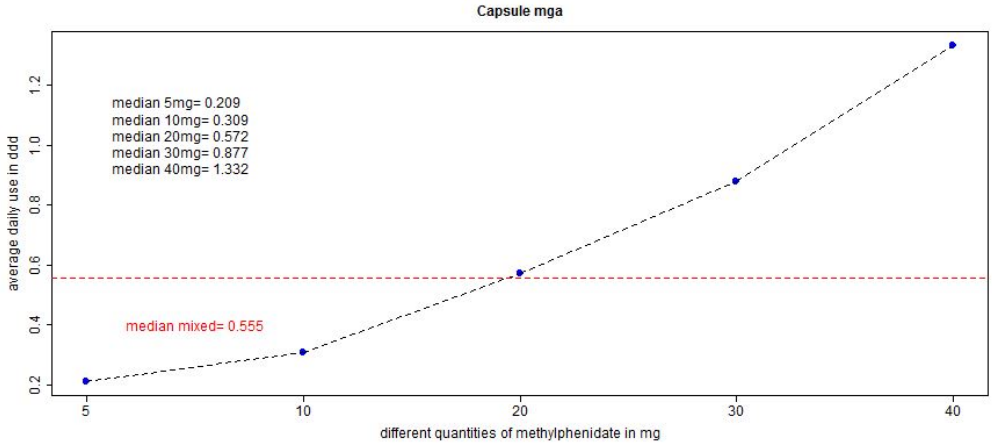


Figure 20: Medians and corresponding CI for methylphenidate capsule

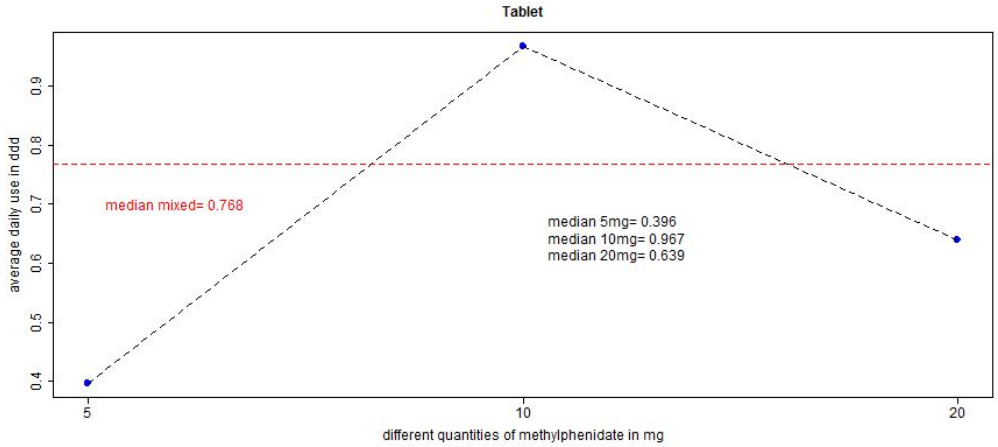


Figure 21: Medians and corresponding CI for methylphenidate tablet

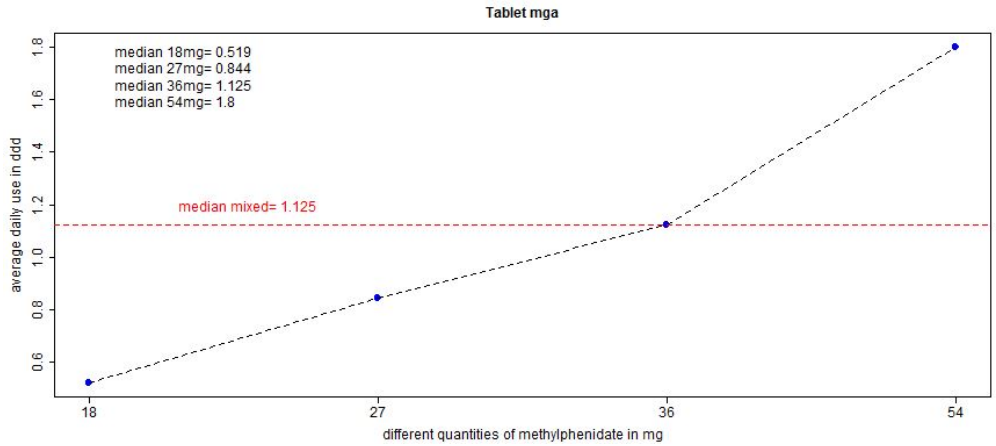


Figure 22: Medians and corresponding CI for methylphenidate tablet mga

Before moving to the next section I would like to focus the attention a bit longer to the values of sildenafil 20 mg. I think it is important to keep in mind that even though we are satisfied with the result obtained so far (the median value which perfectly reflects the expectation based on medical assumptions), there are still some anomalous values which could be a good input for further analyses concerning the use of higher dose than prescribed. Figures 23-25 are based on the average daily use values in ddd of sildenafil 20 mg:

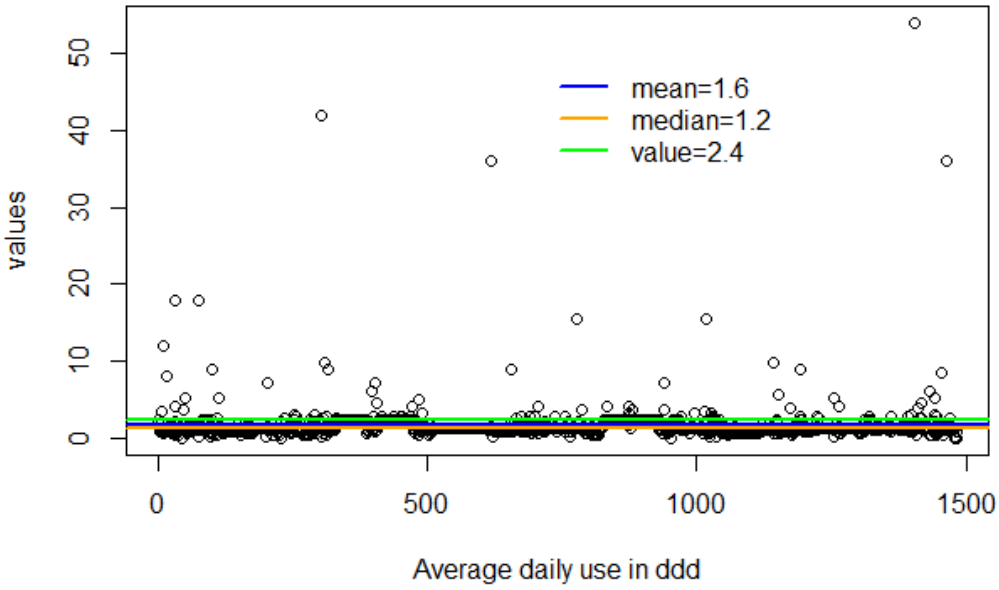


Figure 23: Daily use values of sildenafil 20 mg expressed in ddd

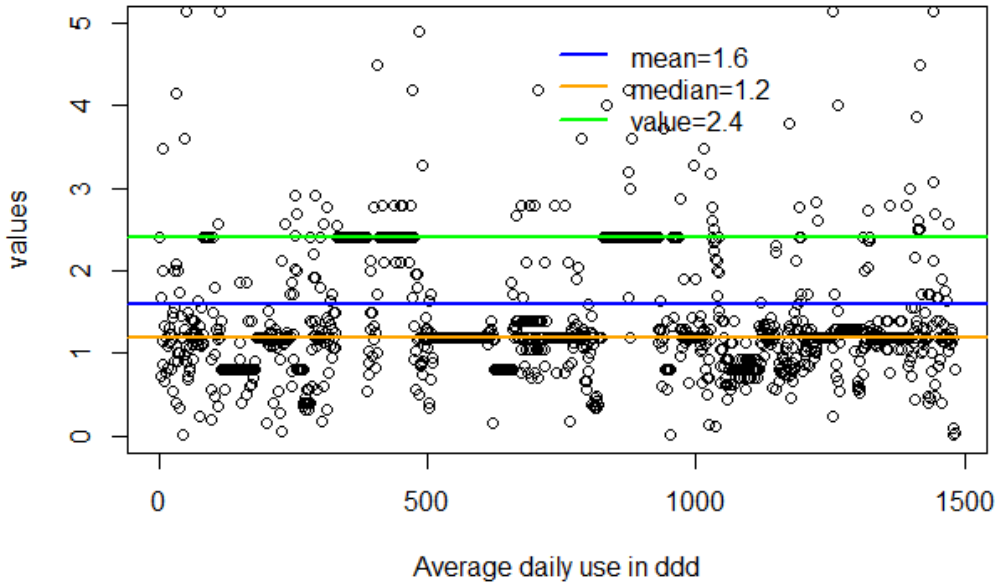


Figure 24: Daily use values of sildenafil 20 mg expressed in ddd (zoomed version)

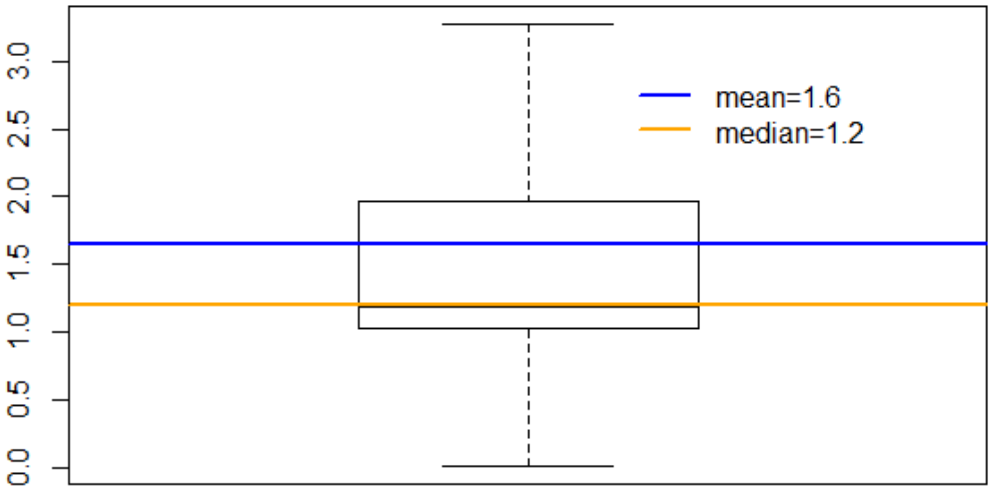


Figure 25: Boxplot average daily use of sildenafil 20 mg expressed in ddd

Figures 23 and 24 show the same data on different y- scale. The points on Figures 23 and 24 are not indicating a person, but the average use expressed in ddd! Each subject can be associated to more than one point: the number of points per subject is strictly related to the number of purchases per subject. The blue line represents the mean (=1.6), the orange line the median (=1.2) and the green line corresponds to a specific numeric value (=2.4). This last one has been drawn because it represents a critical point. In fact, the mean of all the points excluding those greater than 2.4 equals 1.3, while the mean of all the points discarding those greater or equal to 2.4 equals 1.1. By removing all values greater or equal to 2.4, the skewness is also removed and therefore we obtain a mean value which is pretty much close to the median value. The total amount of observations are shown here below:

- 1483 is the total number of sildenafil 20mg average daily use values
- 336 is the total number of points ≥ 2.4, i.e. 22.7%
- 90 is the total number of points > 2.4, i.e. 6.1%

Fig. 26 shows the city contribution to the amount of values greater than the critical point. The percentages in the 2nd and the 4th yellow rows sum up to 100 because they refer to the same total, while the percentages in the 1st and 3rd yellow rows do not sum up to 100, because each individual percentage is derived from a different total. Since most of the data are gathered in Amsterdam and then Utrecht and Eindhoven follow, it is reasonable that Amsterdam is the city that contributes most to the total amount of the "anomalous" values (see the order of the percentages in the 2nd and 4th yellow rows). A huge amount of values coming from Utrecht is bigger or equal to 2.4, about 60% of the total, which means that in Utrecht a considerable group of patients assumes a daily dosage of sildenafil 20mg remarkably higher than the daily defined dose. Unfortunately the available data does not allow any further investigation on this matter.

	AMS	EIN	UTR	TOTAL
Overall number of values	944	314	225	1483
Number of values >=2.4	184	16	136	336
Percentage of the total number of values in the city	19,50%	5,10%	60,40%	
Percentage of the total number of values >=2.4	54,70%	4,80%	40,50%	
Number of values >2.4	54	14	22	90
Percentage of the total number of values in the city	5,70%	4,50%	9,80%	
Percentage of the total number of values >2.4	60%	15,60%	24,40%	

Figure 26: Table with outliers percentages

2.3 Different ways of dealing with missing values

Once the additional variables are created, the data set can be represented as follows:

	ID	Date	City	Dosage	ddd	Dos.in.mg	num.of.tablets	Periods	Avg.daily.use.in.mg	Avg.daily.use.in.ddd
1	0006b0be534ea32c3cd1147512433288	2012-10-02	ams	SILDENAFIL TABLET	50MG	50	8	142	2.8169014	0.056338028
2	0006b0be534ea32c3cd1147512433288	2013-02-21	ams	SILDENAFIL TABLET	50MG	50	8	119	3.3613445	0.067226891
3	0006b0be534ea32c3cd1147512433288	2013-06-20	ams	SILDENAFIL TABLET	50MG	50	8	299	1.3377926	0.026755853
5	0007055afe14a6e07a99d71bdcabc76c	2013-04-10	ams	SILDENAFIL TABLET	100MG	100	4	188	2.1276596	0.042553191
7	000e7d7ffc2d53f5bf0ae060c3696be6	2012-03-13	utr	SILDENAFIL TABLET	50MG	50	10	196	2.5510204	0.051020408
8	000e7d7ffc2d53f5bf0ae060c3696be6	2012-09-25	utr	SILDENAFIL TABLET	50MG	50	5	612	0.4084967	0.008169935

Figure 27: Final data set

The table above is the output of a function that I specifically created in R (one for sildenafil and one for methylphenidate), which can be used for other possible studies of the same type of data. This function

requires as input only a dataset in the same exact format as the one used for the analysis. In case of different variables registered in the dataset or of any tiny misspelling the function will not work, so it is extremely important to carefully check whether the input dataset respects such requirement.

There are several missing (NA) values, all in correspondence to the last period for each subject. Of course, if there is a missing value in the variable "Periods" for a specific subject, then an NA value will consequently appear in all those variables derived from it. Our interest is on the average daily use expressed in ddd, because it is the one from which we want to derive the indicator of central tendency (the median). Imputing methods can be used to deal with missing data. There are several procedures which make different assumptions concerning the nature of the missing values:

1. MCAR = Missing completely at random: data are missing independently of both observed and unobserved data. In this case listwise deletion is a good way to handle missingness (one row is excluded from the study if any single value is missing);
2. MAR = Missing at random: given the observed data, data are missing independently of unobserved data. In this case there are different options for imputing the missing values;
3. MNAR = Missing not at random: missing observations are related to values of unobserved data. In this case is also possible to use different approaches to impute the missing value. In R several libraries which do the same job in diverse ways are available.

Listwise deletion is by far the easiest and the most popular procedure, because of its simplicity. It works rather well when data are missing completely at random: in this situation bias will not be generated when the estimation on the available data is derived and the standard errors will reflect the actual amount of information. It is not really straightforward to conclude which is the missingness nature of the data: the majority of the missing values can be considered as MCAR, with an exception of those patients who purchase sildenafil 20 mg (hence affected by pulmonary hypertension) who can leave the study because of death. For all other patients is difficult to make hypotheses about the missing mechanism.

The problem concerning the sildenafil 20 mg is to determine whether the last medicine purchased has been used by the person. In case of listwise deletion we avoid the problem by not considering the information of the last purchase by removing the corresponding row for every subject. Due to the uncertainty about the nature of the missingness a sensitivity analysis has been performed, which is a technique used to compare different imputation methods (including the listwise deletion). The statistic of interest will be then computed based on the imputed data. The R library `mice` can be used to impute missing data under the assumption of MAR and MNAR.

IMPUTING MISSING VALUES

Let Y_j (where j = 1,. . . ,p) be an incomplete variable among p variables. Denote by Y_j^{mis} and Y_j^{obs} the missing and the observed values respectively. The number of imputations is denoted by $m \geq 1$. The hth imputed dataset is denoted as $Y^{(h)}$ where h = 1,. . . ,m. Let $Y_{-j} = (Y_1, ..., Y_{j-1}, Y_{j+1}, ..., Y_p)$ denote the group of p variables excluded the jth. The mice procedure involves three consequential steps, as depicted from Figure 28.

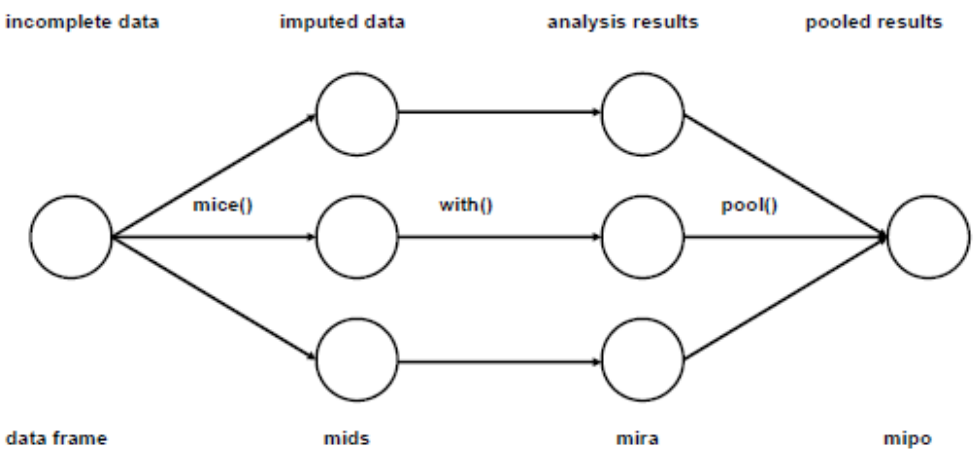


Figure 28: Representation of the main steps for multiple imputation

STEP 1 (first group of arrows): starting from the incomplete dataset, plausible values are drawn from a distribution specifically modeled for each missing entry using the function `mice()`. Suppose Y is a partially observed random sample distributed as a p-variate distribution $P(\vec{Y}|\vec{\theta})$. It is assumed that this distribution depends only on $\vec{\theta}$; the question now is how to derive the parameters of the multivariate distribution of $\vec{\theta}$. The mice algorithm achieves the posterior distribution of θ by sampling iteratively from the following conditional distributions:

$$P(Y_1|Y_{-1}, \theta_1)$$
$$\vdots$$
$$P(Y_p|Y_{-p}, \theta_p)$$

The tth iteration of chained equation is produced by Gibbs sampling that successively draws from the following distributions:

$$\theta_1^{*(t)} \sim P(\theta_1|Y_1^{obs}, Y_2^{t-1}, ..., Y_p^{t-1})$$
$$Y_1^{*(t)} \sim P(Y_1|Y_1^{obs}, Y_2^{t-1}, ..., Y_p^{t-1}, \theta_1^{*(t)})$$
$$\vdots$$
$$\theta_p^{*(t)} \sim P(\theta_p|Y_p^{obs}, Y_2^{t}, ..., Y_{p-1}^{t})$$
$$Y_p^{*(t)} \sim P(Y_p|Y_p^{obs}, Y_2^{t}, ..., Y_p^{t}, \theta_p^{*(t)})$$

In Fig. 28 the numebr of imputation equal 3 is represented, even though is generally preferred m = 5. With this procedure m completed data set are constructed. The imputed datasets are identical to the starting one apart from the missing values, which in the seconds are filled in with some numbers;

STEP 2 (second group of arrows): for each imputed data set the analysis is performed;

STEP 3 (third group of arrows): all m estimated quantities of interest are then pooled together along with their variance [1, Rubin].

On the type of missingness we assume to deal with, either MAR or MNAR. In the latter case, unless we have external data, there is no way to estimate the amount of error. One expedient would be to consider different possible external causes that could either increase or dicrese the estimate of missing values. This can be taken into account by multiplying the imputations by a factor(s) which numerically represent(s) the entity of

the external influence. I found quite hard to determine the sign and size of such a factor, so I created a vector of factors that could cover different possibilities in terms of sign and size. Fig. 29 shows the percentage of missing data for each type of medicine.

TYPE OF MEDICINE	TOTAL	NUMBER OF NA	NUMBER OF OBSERVED	% OF NA
Methylpenidate capsule mga 5 mg	775	127	648	16,39
Methylpenidate capsule mga 10 mg	2362	336	2026	14,23
Methylpenidate capsule mga 20 mg	3668	600	3068	16,36
Methylpenidate capsule mga 30 mg	2661	451	2210	16,95
Methylpenidate capsule mga 40 mg	999	134	865	13,41
Methylphenidate tablet 5 mg	11142	1988	9154	17,84
Methylphenidate tablet 10 mg	47553	7636	39917	16,06
Methylphenidate tablet 20 mg	122	28	94	22,95
Methylphenidate tablet mga 18 mg	5683	731	4952	12,86
Methylphenidate tablet mga 27 mg	5327	822	4505	15,43
Methylphenidate tablet mga 36 mg	13675	1840	11835	13,46
Methylphenidate tablet mga 54 mg	10668	1459	9209	13,68
Sildenafil 20 mg	1638	155	1483	9,46
Sildenafil 25 mg	2019	839	1180	41,56
Sildenafil 50 mg	13251	4827	8424	36,43
Sildenafil 100 mg	17262	5270	11992	30,53

Figure 29: Table with number of missing values (NA)

There is no established cutoff point which determines the maximum amount of missing data allowed for a 'good' analysis. There are some indications ranging from 5% to 20% missing data. As Fig. 29 shows, the percentage of missing data for sildenafil is quite high.

Once the first step in the imputation procedure is performed, the median along with the corresponding variance can be estimated. In this particular case steps 2 and 3 are not necessary. Since the imputation procedure has been done 5 times, 5 different estimates of the median are obtained: $\tilde{\beta} = (\tilde{\beta_1}, ..., \tilde{\beta_5})$. From the 5 imputed datasets we can also derive the 5 *within-imputation variances* by just calculating the variance of the all values of the variable `avg.daily.use.in.ddd`: $\tilde{V} = (\tilde{V_1}, ..., \tilde{V_5})$.

1. MI (Multiple Imputation) estimator of β:

$$\tilde{\beta}_{MI} = \frac{1}{m} \sum_{j=1}^{m} \tilde{\beta}_j$$

2. MI variance estimate

$$\tilde{V}_{MI} = \bar{V} + (1 + M^{-1})B$$

where

$$\bar{V} = \frac{1}{m} \sum_{j=1}^{m} \tilde{V}_i \ (average\ within - imputation\ variance)$$

and

$$B = \frac{1}{m-1} \sum_{j=1}^{m} (\tilde{\beta}_j - \tilde{\beta}_{MI})^2 \ (between\ imputation\ variance)$$

At this stage the main goal is to compare the estimate of the median derived by imputation with the one given by listwise deletion. For this purpose I used the t-statistic. Under the null hypothesis that a single parameter β is zero, it is:

$$\frac{\tilde{\beta} - 0}{\sqrt{\tilde{V}}}$$

on ν degrees of freedom where

$$\nu = (m-1)\left(1+\frac{1}{r}\right)^2$$

where r is the relative increase in variance due to missingness

$$r = \left(1+\frac{1}{m}\right)\frac{B}{\bar{V}}$$

.

In Fig. 30 the median of sildenafil derived with different approaches approaches are shown together with their standard error. For listwise deletion the quantity of interest is calculated by bootstrapping (a procedure that makes inference on the population from a sample by resampling the data and making inference on the new resampled data), appropriate in this case due to the lack of normality. The SE for the two imputing methods is obtained by taking the square root of $\tilde{V}_{MI}$. The standard error is an indication of the reliability of accuracy of the estimated statistics; if it is small it means it is a good reflection of the population median. The SEs for listwise deletion appear all the same due to approximation to 3 digits.

METHOD	MEDIAN 20 mg	MEDIAN 25 mg	MEDIAN 50 mg	MEDIAN 100 mg
Listwise deletion	1,2 (SE: 0,000)	0,059 (SE: 0,002)	0,111 (SE: 0,002)	0,211 (SE: 0,002)
Imputation under MAR	1,2 (SE: 1,005)	0,098 (SE: 1,317)	0,126 (SE: 1,194)	0,195 (SE: 1,469)
Imputation under MNAR	1,2 (SE: 1,201)	0,115 (SE: 1,678)	0,143 (SE: 1,323)	0,217 (SE: 1,661)

Figure 30: Table with imputed medians

The SE for MAR and MNAR are extremely high. Median values derived by imputation are not statistically different from those obtained via listwise deletion. For simplicity the listwise deletion approach is used.

2.4 Creation of the weekly plot

For researcher in this field it is interesting to see a plot of the picture of weekly usage with the median value. This type of plot requires data manipulation. The goal here consists in building a matrix in which every row corresponds to a unique subject and every column represents a single day. All columns represent the total number of days from the first to the last purchase present in the original data set.

Once such a matrix is created, we need to sum up for every row all the values by multiple of 7 (number of days in a week). A new matrix is then created which has the same number of rows as the previous one, but this time with as many columns as the number of weeks forming the overall period of study. From this matrix it is possible to derive the weekly consumption by summing up all the values by columns. To recap, all necessary steps to obtain the weekly use values are described here below:

1. Create a matrix with n rows and k columns, where n is the number of unique subjects of the servey and k the number of days contained in the range of time of the study ("avg" is used as abbreviation for "average daily use in mg", while in parentheses in Fig. 31 the first number is the row index and the second number is the column index).

	DAY 1	DAY 2	...	...	DAY k
SUBJECT 1	avg (1,1)	avg (1,2)			avg (1,k)
SUBJECT 2	avg (2,1)	avg (2,2)			avg (2,k)
...					
...					
...					
SUBJECT n	avg (n,1)	avg (n,2)			avg (n,k)

Figure 31: Table of weekly use - 1° step

In order to fill in the table in Fig. 31 only the complete data are considered. For a missing value a zero is filled in. "DAY 1" is the first day registered in the study and "DAY k" is the last day. For the majority of the subjects "DAY 1" doesn't coincide with the first day they purchased the medicine, consequently most of the first sets of cells in the table have a zero, because in those days preceding the first purchase there is no information about the consumption of the medicinal. Similarly, we cannot precisely estimate in how many days the last quantity of medicine purchased is used by a patient and therefore this piece of information is considered as unknown and represented by a zero in the last cell for a specific person (i.e. all cells corresponding to the last period of a person, which includes all days that run from the date of the last purchase to "DAY k"). Only known and complete information for each person is used to create the plot.

2. For each subject (each row) sum the values by intervals of 7 days.

	WEEK 1 (DAY 1 - DAY 7)	WEEK 2 (DAY 8 - DAY 14)	...	...	WEEK t (DAY k-6 - DAY k)
SUBJECT 1	avg (1,1) + ... + avg (1,7)	avg (1,8) + ... + avg (1,14)			avg (1,k-6) + ... + avg(1,k)
SUBJECT 2	avg (2,1) + ... + avg (2,7)	avg (2,8) + ... + avg (2,14)			avg (2,k-6) + ... + avg (2,k)
...					
...					
...					
SUBJECT n	avg (n,1) + ... + avg (n,7)	avg (n,8) + ... + avg (n,14)			avg (n,k-6) + ... + avg (n,k)

Figure 32: Table of weekly use - 2° step

3. Total weekly use in mg is easily calculated by summing up all the numbers in every column.

	WEEK 1 (DAY 1 - DAY 7)	WEEK 2 (DAY 8 - DAY 14)	...	...	WEEK t (DAY k-6 - DAY k)
SUBJECT 1	avg (1,1) + ... + avg (1,7)	avg (1,8) + ... + avg (1,14)			avg (1,k-6) + ... + avg(1,k)
SUBJECT 2	avg (2,1) + ... + avg (2,7)	avg (2,8) + ... + avg (2,14)			avg (2,k-6) + ... + avg (2,k)
...					
...					
...					
SUBJECT n	avg (n,1) + ... + avg (n,7)	avg (n,8) + ... + avg (n,14)			avg (n,k-6) + ... + avg (n,k)
ALL SUBJECTS	TOTAL WEEK 1 = avg (1,1) + ... + avg (n,7)	TOTAL WEEK 2 = avg (1,8) + ... + avg (n,14)	...	...	TOTAL WEEK t = avg (1, k-6) + ... + avg (n,k)

Figure 33: Table of weekly use - 3° step

In the end all k values are plotted together with the median line, like Fig. 35-36. The median has been chosen as the statistic of interest due to the skewed nature of the weekly use data. In Fig. 34 the weekly usage of sildenafil 25, 50 and 100 mg in all the three cities is illustrated.

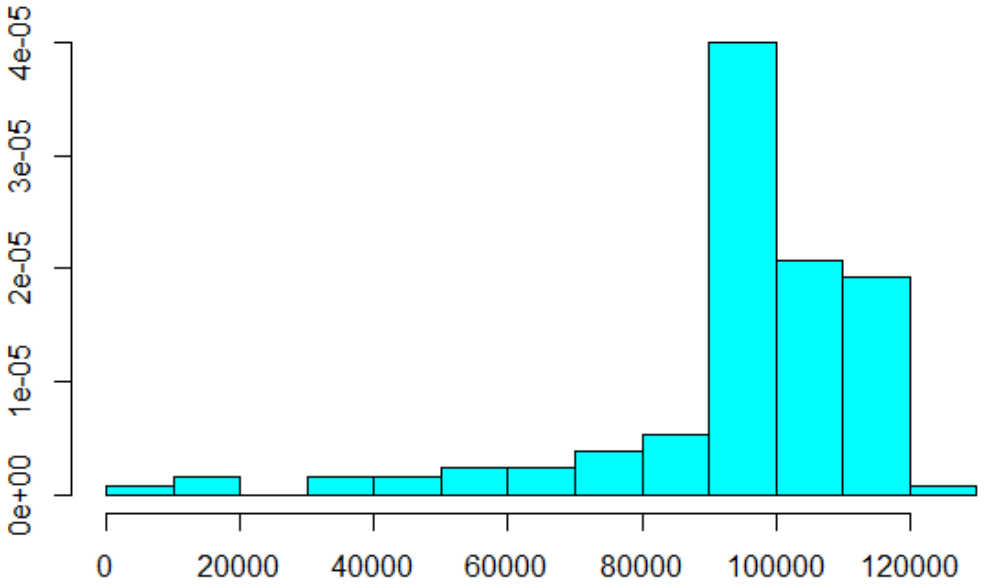


Figure 34: Histogram of sold sildenafil 25, 50, 100 mg per week

As can be seen in Fig. 35-36, there are two different medians: the blue one is calculated considering all k values, while the red one is derived by discarding the first 20 values (in other words it is the median of the k weekly use values excluding the first 20). This distinction made will be explained in the next section. It is important to understand how to read the numbers reported in each graph: the weekly data are given by observed numbers, which perfectly mirror the legal consumption in the cities of Amsterdam, Utrecht and Eindhoven. Any attempt to generalize these results to the overall population of sildenafil and methylphenidate consumers is misleading, since these numbers do not take into account the purchase and use of such medicines in the illegal system.

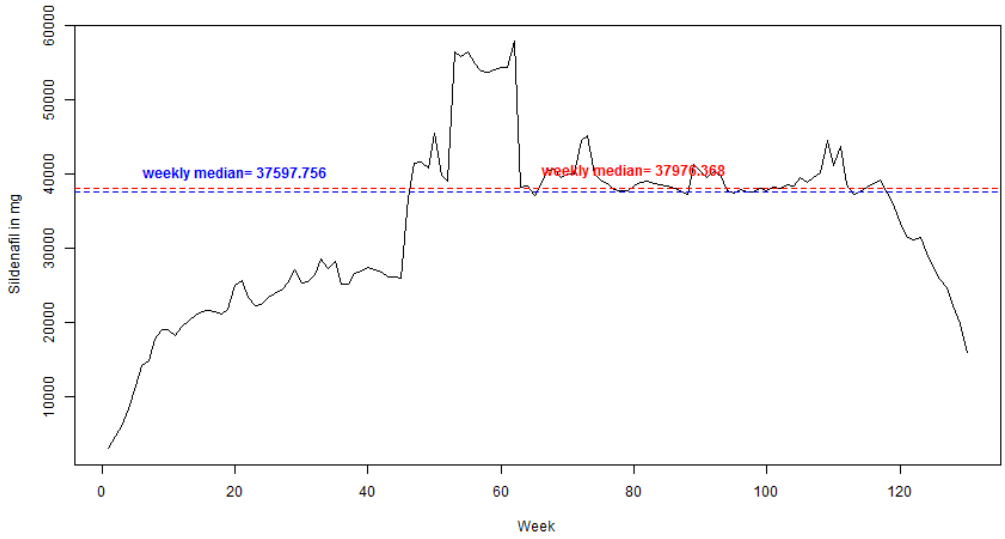


Figure 35: Plot weekly consumption of sildenafil 20 mg

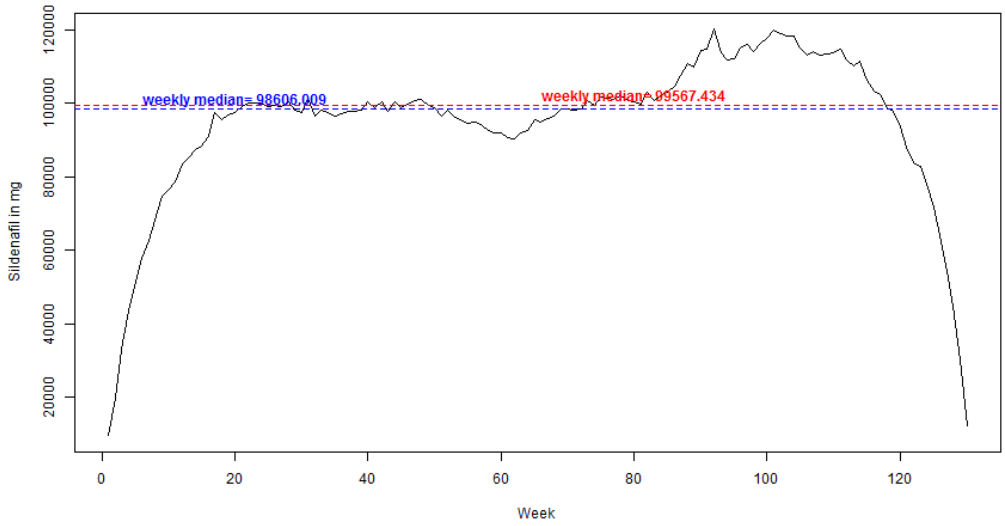


Figure 36: Plot weekly consumption of sildenafil 25, 50, 100 mg

2.5 App construction

Once all the necessary data manipulation were performed and all plots were constructed, it was of interest to gather all of them in a unique interactive system. The R package `shiny` allows to create an application page: by running the R script, a webpage automatically opens in the preset browser. Due to some strong differences between the two types of medicines, two different apps are created: one for sildenafil and one for methylphenidate, even though they look pretty much the same and they contain equivalent plots (more details are given in section 2.3). The first interface that appears by running the R script is the following (sildenafil example):

On the left side of the screen there are four different command options:

- a drop-down list with the possibility to choose between the complete dataset, the data from Amsterdam, Utrecht or Eindhoven;
- choose whether to include the outliers or not in the boxplots;

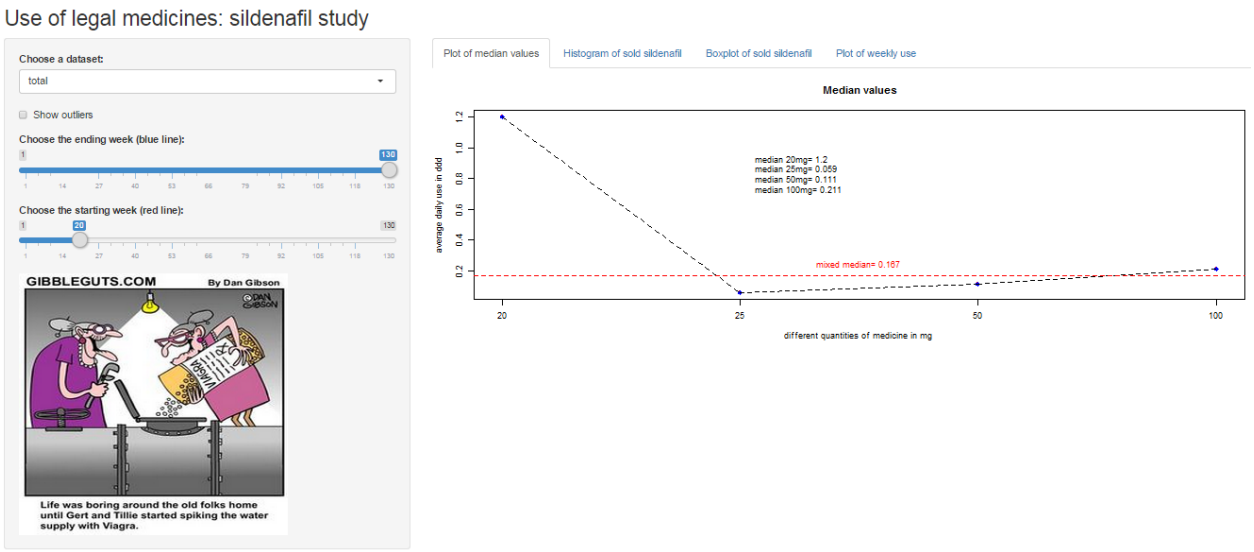


Figure 37: App for sildenafil

- a bar from which to select the number of weeks from the start in order to calculate the median (for example from 0 to 20 weeks, or from 0 to 100 weeks, but always with 0 as starting point);
- a bar from which to select the number of weeks to discard from the start in order to calculate the median (for example from 20 to k weeks, or from 40 to k weeks, but always with k as ending point).

On the right side there are four labels that, once clicked, show the corresponding plot. Going from the left to the right there are:

- the plot with the statistic of interest (the median) distinguished by type of dosage;
- the histogram of sold medicine expressed in ddd;
- the boxplot of sold medicine expressed in ddd;
- the weekly usage plot with the corresponding median value.

Because of the need to distinguish between 20 mg and the group of 25, 50 and 100 mg for sildenafil and between capsule mga, tablet and tablet mga for methylphenidate, a split of the datasets into respectively two and three groups has been necessary before the calculation of the variable "Periods", which could take different values of "DAY 1" and "DAY k" depending on the subset. The two groups of sildenafil have the same first day and last day, but it is not the same for methylphenidate and this can be noticed in the weekly plot (after having run the app), where for tablet category there are 78 weeks available while for capsule mga and tablet mga only 77. After the split and the calculation of "Periods", the subsets have been merged to create a unique dataset again. In the sildenafil app there are 4 (choices from the top-down list) * 2 (split of the dataset) * 4 (types of graphs) = 32 graphs in total.

In the methylphenidate app there are 4 (choices from the top-down list) * 3 (split of the dataset) * 4 (types of graphs) = 48 graphs in total. Due to the large amount of plots available and the fact that all of them can be directly and quickly viewed running the apps. They are not shown in this report. All the pictures reported are those either actually needed for the explanation or not visualizable in the apps.

The library `shiny` requires the creation of at least two `.R` files: the `ui.R` (user interface which establishes the layout of the app) and the `server.R` (which contains the instruction to build the app). There can also be another `.R` file containing additional commands. Such a document is called `script.R` in which the dataset is downloaded and the subdata set according to the city under investigation is created. A function that calculate the cumulative median in the app is defined. Once the two (or more) `.R` files are created, they must

be enclosed in a common folder together with the data required. If one of this document is excluded from the folder, the all procedure for building the app will fail. In order to make the app working it is necessary to open one of the `.R` files in `RStudio` (no matter which one is chosen) and press the `Run App` bottom (which should be available only in the last version of RStudio currently present: `R version 3.1.2 (2014-10-31)`).

2.6 Generalization

The initial plane was to generalize the app to other possible types of medicines and so to be able to produce the same output in terms of plots. This type of interface cannot be created for several reasons:

1. The statistic of interest (in the studied case the median) can easily vary basing on the distribution of the data we deal with. In this case they were skewed, so the median was chosen, but other data could have a symmetric distribution and consequently it would be more appropriate to select the mean. Moreover, since we are dealing with subsets of data and not with the entire data in itself (think about the two groups of revatio and 25, 50, 100 mg for sildenafil) it is not only needed to check the distribution of the overall group of values, but specifically for each subsets, which make the problem even more difficult to face in terms of generalization.

2. Like we have seen for sildenafil and methylphenidate, every type of medicine needs a specific and "personalized" split with regard to the type of dosage. Since I have already known from the beginning which distinction between dosages I had to consider for one or the other medicine, it was "easy" to build the apps in such a way that they could face this division; but, as the result shows, I couldn't find a way to create a unique app that lets decide how many and which splits to make based on the type of data. Furthermore, I cannot foresee which kind of dosages can exist for different medicines and consequently I am not able to make the app capable of recognizing such unknown splits.

3. Description of skills acquired or strengthened

With this internship experience I have had the opportunity to both strengthen my analytical, critical and statistical abilities and to explore and examine in depth some statistical tools. When I started the internship I was faced with a specific problem to solve, which was the result of an analysis previously done. Instead of following the steps of my predecessor, I preferred to start the analysis "from the beginning" being led by what my senses were telling me was the right choice. So partly I took inspiration from my predecessor's work, specifically when I had to deal with data "amplification" (I mean the creation of the new variables), but later I approached the question regarding the statistic of interest in a different way, just asking myself what should I have warned about in order to find a correct statistical measure for the available data.

While analyzing the data I also encountered the problem of missing values. In my limited past experience of data analysis, I always applied the rule of listwise deletion, hence I always considered the complete data. Besides, I was curious to understand if there is a safer or more appropriate way to undertake the issue. For this reason I decided to read papers about the different natures of missingness and multiple imputation. Even though I cannot say I know a lot about this topic, I had the chance to improve my knowledge about it and I gained more awareness on the matter.

Furthermore, the usage of the `shiny` package was something completely unfamiliar to me, even though the courses of the master program made me achieving (I guess) a good ability of programming with the R language. Probably the creation of the app has been the most challenging task to fulfill for the internship project. Thanks to a good guide in internet I could learn how it works and write a proper script for the data; I liked the challenge.

4. Conclusion

Personally I really enjoyed this internship experience, both for the environment and the project. I found the place where I worked really friendly, sharing a room with other students as me and having lunch with all the interns. I especially felt extremely comfortable in my office and even though I didn't manage to speak Dutch, I easily communicated with my roommates, who have been a very good company. I perceived a general atmosphere of calm, serenity and at the same time efficiency and productivity, which I wish could be found in any working environment. Furthermore, the topic of the project was curious and somewhat interesting due to its possible usefulness for the society. In some parts it revealed to be challenging, but I liked the stimulus and I positively caught and faced it. The task gave me also the chance to work independently and so to test my abilities in organizing the activities autonomously and in "looking out for myself". Last but not least, it has been an opportunity to test what it means to work in a real working environment.

5. Bibliography

5.1. Texts and documents

[1] Rubin DB. *Multiple imputation for nonresponse in surveys.* Wiley, 1987.

[2] M.G. Bulmer *Principles of statistics.* Dover, 1979.

[3] M. Pagano and K. Gauvreau *Principles of biostatistics.* Duxbury, 2nd edition, 2000.

[4] S. van Buuren and K. Groothuis-Oudshoorn *mice: Multivariate Imputation by Chained Equations in R.* Journal of Statistical Software, December 2011, Volume 45, Issue 3. (www.jstatsoft.org/v45/i03/paper)

[5] M.G. Kenward and J. Carpenter *Multiple imputation: current perspectives.* Statistical Methods in Medical Research 2007; 16: 199-218.

[6] B. Harding, C. Tremblay and D. Cousineau *Standard errors: A review and evaluation of standard error estimators using Monte Carlo simulations.* The Quantitative Methods for Psychology, 2014, vol. 10, no. 2. (www.tqmp.org/RegularArticles/vol10-2/p107/p107.pdf)

5.2. Websites

[1] *The multiple imputation FAQ page*: sites.stat.psu.edu/~jls/mifaq.html

[2] *Confidence intervals for a single median*:

epilab.ich.ucl.ac.uk/coursematerial/statistics/non_parametric/confidence_interval.html

[3] *DATA MINING Desktop Survival Guide*: datamining.togaware.com/survivor/Mean_Median_Mode.html

[4] *Listwise Deletion: It's NOT Evil*: statisticalhorizons.com/listwise-deletion-its-not-evil

[5] *FAQs - Measures of Central Tendency*: statistics.laerd.com/statistical-guides/measures-central-tendency-mean-mode-median-faqs.php

[6] *WHO Collaborating Centre for Drug Statistics Methodology - Definition and general considerations*: www.whocc.no/ddd/definition_and_general_considera/

5. Appendix: R code

Function to create the final data set for sildenafil.

```
creation_dataset_for_app <- function(input){

  #Reading the data
  data <- read.csv2(input)
  names(data) <- c("ID","Date","City","Dosage","ddd")
  data$ID <- factor(data$ID)
  data$Date <- as.Date(data$Date, "%d-%m-%Y")
  data$City <- factor(data$City)
  data$ddd <- as.numeric(data$ddd)
  data$Dosage <- as.factor(data$Dosage)

  # transforming the variable Dosage in a numeric variable
  Dos.in.mg <- ifelse(data$Dosage=="SILDENAFIL␣SUSP␣ORAAL␣10MG/ML"|data$Dosage=="
      SILDENAFIL␣TABLET␣␣20MG",20,
                      ifelse(data$Dosage=="SILDENAFIL␣TABLET␣␣25MG",25,
                             ifelse(data$Dosage=="SILDENAFIL␣TABLET␣␣50MG",50,100)
                                ))
  d <- data.frame(data,Dos.in.mg)

  # calculation of number of tablets purchased by person
  num.of.tablets <- (50*d$ddd)/d$Dos.in.mg # 50 is the ddd quantity of sildenafil
      expressed in mg
  d1 <- data.frame(d,num.of.tablets)

  # calculation of the periods expressed in days
  periods <- numeric(nrow(data))

  g <- function(data){
    periods <- rep(NA,nrow(data))
    for(i in 1:length(data$Date)){
      if(identical(data$ID[i],data$ID[i+1])==T){
        periods[i] <- as.numeric(difftime(data$Date[i+1],data$Date[i]))
      }
    }
    return(periods)
  }

  Periods <- g(d1)
  a <- data.frame(d1,Periods)

  # Dealing with people who purchase more than one time in the same day
  # Counting the total number of tablets for a person who purchased twice in the
      same day
  for(i in 1:nrow(a)){
    if(identical(a$ID[i],a$ID[i+1])==T){
      if(identical(a$Date[i],a$Date[i+1])==T){
        a$num.of.tablets[i+1] <- a$num.of.tablets[i+1] + a$num.of.tablets[i]
      }
    }
  }

  # Deleting the rows for which the period is equal to 0
  index <- which(a$Periods==0)
  new.data <- a[-index,] # removed the na.omit()
```

```
# For each period the average use per day is calculated
Avg.daily.use.in.mg <- new.data$num.of.tablets*new.data$Dos.in.mg/new.data$
    Periods
Avg.daily.use.in.ddd <- Avg.daily.use.in.mg/50
aa <- data.frame(new.data,Avg.daily.use.in.mg,Avg.daily.use.in.ddd)

# Check the distribution of the average daily use variable
MG20.data <- aa[aa$Dos.in.mg==20,]
MG20 <- MG20.data$Avg.daily.use.in.ddd[is.finite(MG20.data$Avg.daily.use.in.ddd)
    ]
MG25.data <- aa[aa$Dos.in.mg==25,]
MG25 <- MG25.data$Avg.daily.use.in.ddd[is.finite(MG25.data$Avg.daily.use.in.ddd)
    ]
MG50.data <- aa[aa$Dos.in.mg==50,]
MG50 <- MG50.data$Avg.daily.use.in.ddd[is.finite(MG50.data$Avg.daily.use.in.ddd)
    ]
MG100.data <- aa[aa$Dos.in.mg==100,]
MG100 <- MG100.data$Avg.daily.use.in.ddd[is.finite(MG100.data$Avg.daily.use.in.
    ddd)]

library(MASS)
truehist(MG20)
truehist(MG25)
truehist(MG50)
truehist(MG100)
truehist(aa$Avg.daily.use.in.ddd)

Else.data <- aa[aa$Dos.in.mg!=20,]

##### FIRST PART OF THE COMPLETE DATASET (REVATIO)
data <- MG20.data
data$ID <- factor(data$ID)
data$Date <- as.Date(data$Date)
data$City <- factor(data$City)
data$ddd <- as.numeric(data$ddd)
data$Dosage <- as.factor(data$Dosage)
index <- which(data$Periods==0)
Data <- data
first_day <- min(Data$Date)
last_day <- max(Data$Date)
days <- as.numeric(last_day-first_day)

# ID
library(plyr)
counts <- count(Data,"ID")[,2]
ID <- rep(unique(Data$ID),counts+1)

# Date
chunks <- split(Data$Date,Data$ID)
date <- lapply(chunks,function(x) {
  c(as.Date(first_day),as.Date(x))
})

Date <- unname(do.call("c",date))

d <- data.frame(ID,Date)

# Periods
g <- function(data){
```

```
  periods <- rep(NA,nrow(data))
  for(i in 1:length(data$Date)){
    if(identical(data$ID[i],data$ID[i+1])==T){
      periods[i] <- as.numeric(difftime(data$Date[i+1],data$Date[i]))
    }
  }
  return(periods)
}

Periods <- g(d)

# Avg.daily.use.in.mg
chunks2 <- split(Data$Avg.daily.use.in.mg,Data$ID)
avg.daily.use.in.mg <- lapply(chunks2,function(x){
  c(0,x)
})

Avg.daily.use.in.mg <- unname(unlist(avg.daily.use.in.mg))

# City
chunks3 <- split(Data$City,Data$ID)
city <- lapply(chunks3,function(x){
  c(NA,as.character(x))
})

City <- unname(do.call("c",city))
vector.city <- as.character(tapply(as.character(Data$City),Data$ID,unique))
City[is.na(City)] <- vector.city

# Dosage in mg
chunks4 <- split(Data$Dos.in.mg,Data$ID)
dos.in.mg <- lapply(chunks4,function(x){
  c(20,x)
})

Dos.in.mg <- unname(do.call("c",dos.in.mg))


# Avg.daily.use.in.ddd
chunks5 <- split(Data$Avg.daily.use.in.ddd,Data$ID)
avg.daily.use.in.ddd <- lapply(chunks5,function(x){
  c(0,x)
})

Avg.daily.use.in.ddd <- unname(unlist(avg.daily.use.in.ddd))

# ddd
chunks6 <- split(Data$ddd,Data$ID)
d <- lapply(chunks6,function(x){
  c(0,x)
})
ddd <- unname(do.call("c",d))

D <- data.frame(ID,Date,City,Periods,Dos.in.mg,ddd,Avg.daily.use.in.mg,Avg.daily
    .use.in.ddd)
index <- which(D$Periods==0)
DD <- D[-index,]

med1 <- median(na.omit(DD$Avg.daily.use.in.mg))
```

```
med2 <- median(na.omit(DD$Avg.daily.use.in.ddd))

DD$Avg.daily.use.in.mg[is.na(DD$Avg.daily.use.in.mg)==T] <- 0
DD$Avg.daily.use.in.ddd[is.na(DD$Avg.daily.use.in.ddd)==T] <- 0

for(i in 1:nrow(DD)){
  if(is.na(DD$Periods[i])==T){
    DD$Periods[i] <- as.numeric(last_day-DD$Date[i])
  }
}

###### SECOND PART OF THE COMPLETE DATASET (ERECTILE DYSFUNCTION)
data <- Else.data
data$ID <- factor(data$ID)
data$Date <- as.Date(data$Date)
data$City <- factor(data$City)
data$ddd <- as.numeric(data$ddd)
data$Dosage <- as.factor(data$Dosage)
Data <- data
first_day <- min(Data$Date)
last_day <- max(Data$Date)
days <- as.numeric(last_day-first_day)

# ID
library(plyr)
counts <- count(Data,"ID")[,2]
ID <- rep(unique(Data$ID),counts+1)

# Date
chunks <- split(Data$Date,Data$ID)
date <- lapply(chunks,function(x) {
  c(as.Date(first_day),as.Date(x))
})

Date <- unname(do.call("c",date))

d <- data.frame(ID,Date)

# Periods
g <- function(data){
  periods <- rep(NA,nrow(data))
  for(i in 1:length(data$Date)){
    if(identical(data$ID[i],data$ID[i+1])==T){
      periods[i] <- as.numeric(difftime(data$Date[i+1],data$Date[i]))
    }
  }
  return(periods)
}

Periods <- g(d)

# Avg.daily.use.in.mg
chunks2 <- split(Data$Avg.daily.use.in.mg,Data$ID)
avg.daily.use.in.mg <- lapply(chunks2,function(x){
  c(0,x)
})

Avg.daily.use.in.mg <- unname(unlist(avg.daily.use.in.mg))
```

```
  # City
  chunks3 <- split(Data$City,Data$ID)
  city <- lapply(chunks3,function(x){
    c(NA,as.character(x))
  })

  City <- unname(do.call("c",city))
  vector.city <- as.character(tapply(as.character(Data$City),Data$ID,unique))
  City[is.na(City)] <- vector.city

  # Dosage in mg
  chunks4 <- split(Data$Dos.in.mg,Data$ID)
  dos.in.mg <- lapply(chunks4,function(x){
    c(NA,x)
  })

  Dos.in.mg <- unname(do.call("c",dos.in.mg))

  # Avg.daily.use.in.ddd
  chunks5 <- split(Data$Avg.daily.use.in.ddd,Data$ID)
  avg.daily.use.in.ddd <- lapply(chunks5,function(x){
    c(0,x)
  })

  Avg.daily.use.in.ddd <- unname(unlist(avg.daily.use.in.ddd))

  # ddd
  chunks6 <- split(Data$ddd,Data$ID)
  d <- lapply(chunks6,function(x){
    c(NA,x)
  })
  ddd <- unname(do.call("c",d))

  D2 <- data.frame(ID,Date,City,Periods,Dos.in.mg,ddd,Avg.daily.use.in.mg,Avg.
      daily.use.in.ddd)
  index <- which(D2$Periods==0)
  DD2 <- D2[-index,]

  med1 <- median(na.omit(DD2$Avg.daily.use.in.mg))
  med2 <- median(na.omit(DD2$Avg.daily.use.in.ddd))

  DD2$Avg.daily.use.in.mg[is.na(DD2$Avg.daily.use.in.mg)==T] <- 0
  DD2$Avg.daily.use.in.ddd[is.na(DD2$Avg.daily.use.in.ddd)==T] <- 0

  for(i in 1:nrow(DD2)){
    if(is.na(DD2$Periods[i])==T){
      DD2$Periods[i] <- as.numeric(last_day-DD2$Date[i])
    }
  }

  sildenafil <- rbind(DD,DD2)
  # Save the final dataset
  write.csv(sildenafil,file="sildenafil.csv")

}

# Use the function by writing the name of the dataset to transform in parentheses
creation_dataset_for_app('sales_data␣2012-2014.csv')
```

Function to create the final data set for methylphenidate.

```
creation_dataset_for_app <- function(input){

  #Reading the data
  data <- read.csv2(input)
  names(data) <- c("ID","Date","City","Dosage","ddd")
  data$ID <- factor(data$ID)
  data$Date <- as.Date(data$Date, "%d-%m-%Y")
  data$City <- factor(data$City)
  data$ddd <- as.numeric(data$ddd)
  data$Dosage <- as.factor(data$Dosage)

  # transforming the variable Dosage in a numeric variable
  Dos.in.mg <- ifelse(data$Dosage=="SILDENAFIL␣SUSP␣ORAAL␣10MG/ML"|data$Dosage=="
      SILDENAFIL␣TABLET␣20MG",20,
                ifelse(data$Dosage=="SILDENAFIL␣TABLET␣FO␣␣25MG",25,
                ifelse(data$Dosage=="SILDENAFIL␣TABLET␣FO␣␣50MG",50,
                ifelse(data$Dosage=="SILDENAFIL␣TABLET␣FO␣100MG",100,
                ifelse(data$Dosage=="METHYLFENIDAAT␣CAPSULE␣MGA␣␣5MG"|data$Dosage==
                    "METHYLFENIDAAT␣TABLET␣␣5MG",5,
                ifelse(data$Dosage=="METHYLFENIDAAT␣CAPSULE␣MGA␣10MG"|data$Dosage==
                    "METHYLFENIDAAT␣TABLET␣10MG",10,
                ifelse(data$Dosage=="METHYLFENIDAAT␣CAPSULE␣MGA␣20MG"|data$Dosage==
                    "METHYLFENIDAAT␣TABLET␣20MG",20,
                ifelse(data$Dosage=="METHYLFENIDAAT␣CAPSULE␣MGA␣30MG",30,
                ifelse(data$Dosage=="METHYLFENIDAAT␣CAPSULE␣MGA␣40MG",40,
                ifelse(data$Dosage=="METHYLFENIDAAT␣TABLET␣MGA␣18MG",18,
                ifelse(data$Dosage=="METHYLFENIDAAT␣TABLET␣MGA␣27MG",27,
                ifelse(data$Dosage=="METHYLFENIDAAT␣TABLET␣MGA␣36MG",36,54)))))))))
                    )))
  d <- data.frame(data,Dos.in.mg)

  # calculation of number of tablets purchased by person
  num.of.tablets <- numeric(nrow(d))
  for(i in 1:nrow(d)){
    if(d$Dosage[i]=="SILDENAFIL␣SUSP␣ORAAL␣10MG/ML"|d$Dosage[i]=="SILDENAFIL␣
        TABLET␣20MG"|
          d$Dosage[i]=="SILDENAFIL␣TABLET␣FO␣␣25MG"|d$Dosage[i]=="SILDENAFIL␣TABLET
              ␣FO␣␣50MG"|
          d$Dosage[i]=="SILDENAFIL␣TABLET␣FO␣100MG"){
      num.of.tablets[i] <- (50*d$ddd[i])/d$Dos.in.mg[i] # 50 is the ddd quantity
          of sildenafil expressed in mg
    } else {
      num.of.tablets[i] <- (30*d$ddd[i])/d$Dos.in.mg[i] # 30 is the ddd quantity
          of methylphenidate expressed
    }
  }

  d1 <- data.frame(d,num.of.tablets)

  # calculation of the periods expressed in days
  periods <- numeric(nrow(data))

  g <- function(data){
    periods <- rep(NA,nrow(data))
    for(i in 1:length(data$Date)){
      if(identical(data$ID[i],data$ID[i+1])==T){
```

```
      periods[i] <- as.numeric(difftime(data$Date[i+1],data$Date[i]))
    }
  }
  return(periods)
}

Periods <- g(d1)
a <- data.frame(d1,Periods)

# Dealing with people who purchase more than one time in the same day
# Counting the total number of tablets for a person who purchased twice in the
    same day
for(i in 1:nrow(a)){
  if(identical(a$ID[i],a$ID[i+1])==T){
    if(identical(a$Date[i],a$Date[i+1])==T){
      a$num.of.tablets[i+1] <- a$num.of.tablets[i+1] + a$num.of.tablets[i]
    }
  }
}

# Deleting the rows for which the period is equal to 0 (double purchase in the
    same day)
index <- which(a$Periods==0)
new.data <- a[-index,]

# for each period the average use per day is calculated
Avg.daily.use.in.mg <- new.data$num.of.tablets*new.data$Dos.in.mg/new.data$
    Periods

Avg.daily.use.in.ddd <- numeric(nrow(new.data))
for(i in 1:nrow(new.data)){
  if(new.data$Dosage[i]=="SILDENAFIL␣SUSP␣ORAAL␣10MG/ML"|new.data$Dosage[i]=="
      SILDENAFIL␣TABLET␣20MG"|
        new.data$Dosage[i]=="SILDENAFIL␣TABLET␣FO␣␣25MG"|new.data$Dosage[i]=="
            SILDENAFIL␣TABLET␣FO␣␣50MG"|
        new.data$Dosage[i]=="SILDENAFIL␣TABLET␣FO␣100MG"){
    Avg.daily.use.in.ddd[i] <- Avg.daily.use.in.mg[i]/50
  } else {
    Avg.daily.use.in.ddd[i] <- Avg.daily.use.in.mg[i]/30
  }
}

aa <- data.frame(new.data,Avg.daily.use.in.mg,Avg.daily.use.in.ddd)

# Check the distribution of the average daily use variable
cap.mga.M5.data <- aa[aa$Dosage=="METHYLFENIDAAT␣CAPSULE␣MGA␣␣5MG",]
cap.mga.M5 <- cap.mga.M5.data$Avg.daily.use.in.ddd
cap.mga.M10.data <- aa[aa$Dosage=="METHYLFENIDAAT␣CAPSULE␣MGA␣10MG",]
cap.mga.M10 <- cap.mga.M10.data$Avg.daily.use.in.ddd
cap.mga.M20.data <- aa[aa$Dosage=="METHYLFENIDAAT␣CAPSULE␣MGA␣20MG",]
cap.mga.M20 <- cap.mga.M20.data$Avg.daily.use.in.ddd
cap.mga.M30.data <- aa[aa$Dosage=="METHYLFENIDAAT␣CAPSULE␣MGA␣30MG",]
cap.mga.M30 <- cap.mga.M30.data$Avg.daily.use.in.ddd
cap.mga.M40.data <- aa[aa$Dosage=="METHYLFENIDAAT␣CAPSULE␣MGA␣40MG",]
cap.mga.M40 <- cap.mga.M40.data$Avg.daily.use.in.ddd
tab.M5.data <- aa[aa$Dosage=="METHYLFENIDAAT␣TABLET␣␣5MG",]
tab.M5 <- tab.M5.data$Avg.daily.use.in.ddd
tab.M10.data <- aa[aa$Dosage=="METHYLFENIDAAT␣TABLET␣10MG",]
tab.M10 <- tab.M10.data$Avg.daily.use.in.ddd
```

```
tab.M20.data <- aa[aa$Dosage=="METHYLFENIDAAT␣TABLET␣20MG",]
tab.M20 <- tab.M20.data$Avg.daily.use.in.ddd
tab.mga.M18.data <- aa[aa$Dosage=="METHYLFENIDAAT␣TABLET␣MGA␣18MG",]
tab.mga.M18 <- tab.mga.M18.data$Avg.daily.use.in.ddd
tab.mga.M27.data <- aa[aa$Dosage=="METHYLFENIDAAT␣TABLET␣MGA␣27MG",]
tab.mga.M27 <- tab.mga.M27.data$Avg.daily.use.in.ddd
tab.mga.M36.data <- aa[aa$Dosage=="METHYLFENIDAAT␣TABLET␣MGA␣36MG",]
tab.mga.M36 <- tab.mga.M36.data$Avg.daily.use.in.ddd
tab.mga.M54.data <- aa[aa$Dosage=="METHYLFENIDAAT␣TABLET␣MGA␣54MG",]
tab.mga.M54 <- tab.mga.M54.data$Avg.daily.use.in.ddd

library(MASS)
truehist(cap.mga.M5)
truehist(cap.mga.M10)
truehist(cap.mga.M20)
truehist(cap.mga.M30)
truehist(cap.mga.M40)
truehist(tab.M5)
truehist(tab.M10)
truehist(tab.M20)
truehist(tab.mga.M18)
truehist(tab.mga.M27)
truehist(tab.mga.M36)
truehist(tab.mga.M54)

capsule.mga <- aa[aa$Dosage=="METHYLFENIDAAT␣CAPSULE␣MGA␣␣5MG"|aa$Dosage=="
    METHYLFENIDAAT␣CAPSULE␣MGA␣10MG"|aa$Dosage=="METHYLFENIDAAT␣CAPSULE␣MGA␣20MG"
    |aa$Dosage=="METHYLFENIDAAT␣CAPSULE␣MGA␣30MG"|aa$Dosage=="METHYLFENIDAAT␣
    CAPSULE␣MGA␣40MG",]
tablet <- aa[aa$Dosage=="METHYLFENIDAAT␣TABLET␣␣5MG"|aa$Dosage=="METHYLFENIDAAT␣
    TABLET␣10MG"|aa$Dosage=="METHYLFENIDAAT␣TABLET␣20MG",]
tablet.mga <- aa[aa$Dosage=="METHYLFENIDAAT␣TABLET␣MGA␣18MG"|aa$Dosage=="
    METHYLFENIDAAT␣TABLET␣MGA␣27MG"|aa$Dosage=="METHYLFENIDAAT␣TABLET␣MGA␣36MG"|
    aa$Dosage=="METHYLFENIDAAT␣TABLET␣MGA␣54MG",]

##### FIRST PART OF THE DATASET (METHYLPHENIDATE CAPSULE MGA)
data <- capsule.mga
data$ID <- factor(data$ID)
data$Date <- as.Date(data$Date)
data$City <- factor(data$City)
data$ddd <- as.numeric(data$ddd)
data$Dosage <- as.factor(data$Dosage)
Data <- data
first_day <- min(Data$Date)
last_day <- max(Data$Date)
days <- as.numeric(last_day-first_day)

# ID
library(plyr)
counts <- count(Data,"ID")[,2]
ID <- rep(unique(Data$ID),counts+1)

# Date
chunks <- split(Data$Date,Data$ID)
date <- lapply(chunks,function(x) {
  c(as.Date(first_day),as.Date(x))
})

Date <- unname(do.call("c",date))
```

```
d <- data.frame(ID,Date)

# Periods
g <- function(data){
  periods <- rep(NA,nrow(data))
  for(i in 1:length(data$Date)){
    if(identical(data$ID[i],data$ID[i+1])==T){
      periods[i] <- as.numeric(difftime(data$Date[i+1],data$Date[i]))
    }
  }
  return(periods)
}

Periods <- g(d)

# Avg.daily.use.in.mg
chunks2 <- split(Data$Avg.daily.use.in.mg,Data$ID)
avg.daily.use.in.mg <- lapply(chunks2,function(x){
  c(0,x)
})

Avg.daily.use.in.mg <- unname(unlist(avg.daily.use.in.mg))

# City
chunks3 <- split(Data$City,Data$ID)
city <- lapply(chunks3,function(x){
  c(NA,as.character(x))
})

City <- unname(do.call("c",city))
vector.city <- as.character(tapply(as.character(Data$City),Data$ID,unique))
City[is.na(City)] <- vector.city

# Dosage in mg
chunks4 <- split(Data$Dos.in.mg,Data$ID)
dos.in.mg <- lapply(chunks4,function(x){
  c("cap",x)
})

Dos.in.mg <- unname(do.call("c",dos.in.mg))


# Avg.daily.use.in.ddd
chunks5 <- split(Data$Avg.daily.use.in.ddd,Data$ID)
avg.daily.use.in.ddd <- lapply(chunks5,function(x){
  c(0,x)
})

Avg.daily.use.in.ddd <- unname(unlist(avg.daily.use.in.ddd))

# ddd
chunks6 <- split(Data$ddd,Data$ID)
d <- lapply(chunks6,function(x){
  c(0,x)
})
ddd <- unname(do.call("c",d))

# Dosage
```

```
chunks7 <- split(Data$Dosage,Data$ID)
dosage <- lapply(chunks7,function(x){
  c("cap",as.character(x))
})

Dosage <- unname(do.call("c",dosage))

D <- data.frame(ID,Date,City,Periods,Dosage,Dos.in.mg,ddd,Avg.daily.use.in.mg,
    Avg.daily.use.in.ddd)
index <- which(D$Periods==0)
DD <- D[-index,]

med1 <- median(na.omit(DD$Avg.daily.use.in.mg))
med2 <- median(na.omit(DD$Avg.daily.use.in.ddd))

DD$Avg.daily.use.in.mg[is.na(DD$Avg.daily.use.in.mg)==T] <- 0
DD$Avg.daily.use.in.ddd[is.na(DD$Avg.daily.use.in.ddd)==T] <- 0

for(i in 1:nrow(DD)){
  if(is.na(DD$Periods[i])==T){
    DD$Periods[i] <- as.numeric(last_day-DD$Date[i])
  }
}

##### SECOND PART OF THE DATASET (METHYLPHENIDATE TABLET)
data <- tablet
data$ID <- factor(data$ID)
data$Date <- as.Date(data$Date)
data$City <- factor(data$City)
data$ddd <- as.numeric(data$ddd)
data$Dosage <- as.factor(data$Dosage)

index <- which(data$Periods==0)
Data <- data
first_day <- min(Data$Date)
last_day <- max(Data$Date)
days <- as.numeric(last_day-first_day)

# ID
library(plyr)
counts <- count(Data,"ID")[,2]
ID <- rep(unique(Data$ID),counts+1)

# Date
chunks <- split(Data$Date,Data$ID)
date <- lapply(chunks,function(x) {
  c(as.Date(first_day),as.Date(x))
})

Date <- unname(do.call("c",date))

d <- data.frame(ID,Date)

# Periods
g <- function(data){
  periods <- rep(NA,nrow(data))
  for(i in 1:length(data$Date)){
    if(identical(data$ID[i],data$ID[i+1])==T){
      periods[i] <- as.numeric(difftime(data$Date[i+1],data$Date[i]))
```

```
    }
  }
  return(periods)
}

Periods <- g(d)

# Avg.daily.use.in.mg
chunks2 <- split(Data$Avg.daily.use.in.mg,Data$ID)
avg.daily.use.in.mg <- lapply(chunks2,function(x){
  c(0,x)
})

Avg.daily.use.in.mg <- unname(unlist(avg.daily.use.in.mg))

# City
chunks3 <- split(Data$City,Data$ID)
city <- lapply(chunks3,function(x){
  c(NA,as.character(x))
})

City <- unname(do.call("c",city))
vector.city <- as.character(tapply(as.character(Data$City),Data$ID,unique))
City[is.na(City)] <- vector.city

# Dosage in mg
chunks4 <- split(Data$Dos.in.mg,Data$ID)
dos.in.mg <- lapply(chunks4,function(x){
  c("tab",x)
})

Dos.in.mg <- unname(do.call("c",dos.in.mg))


# Avg.daily.use.in.ddd
chunks5 <- split(Data$Avg.daily.use.in.ddd,Data$ID)
avg.daily.use.in.ddd <- lapply(chunks5,function(x){
  c(0,x)
})

Avg.daily.use.in.ddd <- unname(unlist(avg.daily.use.in.ddd))

# ddd
chunks6 <- split(Data$ddd,Data$ID)
d <- lapply(chunks6,function(x){
  c(NA,x)
})
ddd <- unname(do.call("c",d))

# Dosage
chunks7 <- split(Data$Dosage,Data$ID)
dosage <- lapply(chunks7,function(x){
  c("tab",as.character(x))
})

Dosage <- unname(do.call("c",dosage))

D2 <- data.frame(ID,Date,City,Periods,Dosage,Dos.in.mg,ddd,Avg.daily.use.in.mg,
    Avg.daily.use.in.ddd)
```

```
index <- which(D2$Periods==0)
DD2 <- D2[-index,]

med1 <- median(na.omit(DD2$Avg.daily.use.in.mg))
med2 <- median(na.omit(DD2$Avg.daily.use.in.ddd))

DD2$Avg.daily.use.in.mg[is.na(DD2$Avg.daily.use.in.mg)==T] <- 0
DD2$Avg.daily.use.in.ddd[is.na(DD2$Avg.daily.use.in.ddd)==T] <- 0

for(i in 1:nrow(DD2)){
  if(is.na(DD2$Periods[i])==T){
    DD2$Periods[i] <- as.numeric(last_day-DD2$Date[i])
  }
}

##### THIRD PART OF THE DATASET (METHYLPHENIDATE TABLET MGA)
data <- tablet.mga
data$ID <- factor(data$ID)
data$Date <- as.Date(data$Date)
data$City <- factor(data$City)
data$ddd <- as.numeric(data$ddd)
data$Dosage <- as.factor(data$Dosage)

index <- which(data$Periods==0)
Data <- data
first_day <- min(Data$Date)
last_day <- max(Data$Date)
days <- as.numeric(last_day-first_day)

# ID
library(plyr)
counts <- count(Data,"ID")[,2]
ID <- rep(unique(Data$ID),counts+1)

# Date
chunks <- split(Data$Date,Data$ID)
date <- lapply(chunks,function(x) {
  c(as.Date(first_day),as.Date(x))
})

Date <- unname(do.call("c",date))

d <- data.frame(ID,Date)

# Periods
g <- function(data){
  periods <- rep(NA,nrow(data))
  for(i in 1:length(data$Date)){
    if(identical(data$ID[i],data$ID[i+1])==T){
      periods[i] <- as.numeric(difftime(data$Date[i+1],data$Date[i]))
    }
  }
  return(periods)
}

Periods <- g(d)

# Avg.daily.use.in.mg
chunks2 <- split(Data$Avg.daily.use.in.mg,Data$ID)
```

```
avg.daily.use.in.mg <- lapply(chunks2,function(x){
  c(0,x)
})

Avg.daily.use.in.mg <- unname(unlist(avg.daily.use.in.mg))

# City
chunks3 <- split(Data$City,Data$ID)
city <- lapply(chunks3,function(x){
  c(NA,as.character(x))
})

City <- unname(do.call("c",city))
vector.city <- as.character(tapply(as.character(Data$City),Data$ID,unique))
City[is.na(City)] <- vector.city

# Dosage in mg
chunks4 <- split(Data$Dos.in.mg,Data$ID)
dos.in.mg <- lapply(chunks4,function(x){
  c("tab␣mga",x)
})

Dos.in.mg <- unname(do.call("c",dos.in.mg))


# Avg.daily.use.in.ddd
chunks5 <- split(Data$Avg.daily.use.in.ddd,Data$ID)
avg.daily.use.in.ddd <- lapply(chunks5,function(x){
  c(0,x)
})

Avg.daily.use.in.ddd <- unname(unlist(avg.daily.use.in.ddd))

# ddd
chunks6 <- split(Data$ddd,Data$ID)
d <- lapply(chunks6,function(x){
  c(NA,x)
})
ddd <- unname(do.call("c",d))

# Dosage
chunks7 <- split(Data$Dosage,Data$ID)
dosage <- lapply(chunks7,function(x){
  c("tab_mga",as.character(x))
})

Dosage <- unname(do.call("c",dosage))

D3 <- data.frame(ID,Date,City,Periods,Dosage,Dos.in.mg,ddd,Avg.daily.use.in.mg,
    Avg.daily.use.in.ddd)
index <- which(D3$Periods==0)
DD3 <- D3[-index,]

med1 <- median(na.omit(DD3$Avg.daily.use.in.mg))
med2 <- median(na.omit(DD3$Avg.daily.use.in.ddd))

DD3$Avg.daily.use.in.mg[is.na(DD3$Avg.daily.use.in.mg)==T] <- 0
DD3$Avg.daily.use.in.ddd[is.na(DD3$Avg.daily.use.in.ddd)==T] <- 0
```

```
  for(i in 1:nrow(DD3)){
    if(is.na(DD3$Periods[i])==T){
      DD3$Periods[i] <- as.numeric(last_day-DD3$Date[i])
    }
  }


  methylphenidate <- rbind(DD,DD2,DD3)
  write.csv(methylphenidate,file="methylphenidate.csv")

}

# The only thing is needed to be changed is the input name to give to the function
creation_dataset_for_app('sales_data␣2013.csv')
```

R code to create the app for sildenafil.

User Interface (ui.R)

```
# sildenafil App

shinyUI(fluidPage(

  titlePanel("Use of legal medicines: sildenafil study"),

  sidebarLayout(
    sidebarPanel(
      selectInput("dataset", "Choose a dataset:",
                  choices = c("total", "ams", "utr", "ein")),

      checkboxInput("outliers", "Show outliers", FALSE),

      sliderInput("weeks", "Choose the ending week (blue line):",
                  min=1, max=130, value=130),

      sliderInput("weeks.inv", "Choose the starting week (red line):",
                  min=1, max=130, value=20),

      img(src="e9a0176beb4696dee9828d8e4c14124e.jpg",height=370,width=380)
    ),

    mainPanel(
      tabsetPanel(type = "tabs",
                  tabPanel("Plot of median values", plotOutput("medianPlot")),
                  tabPanel("Histogram of sold sildenafil", plotOutput("histogram",
                      height = "1000px")),
                  tabPanel("Boxplot of sold sildenafil", plotOutput("boxplot",
                      height="800px")),
                  tabPanel("Plot of weekly use", plotOutput("weekly.use.plot",
                      height = "1000px"))
      )
    )
  )
))
```

Server (server.R)

```
# sildenafil App

source("script.R")

shinyServer(function(input, output) {

  datasetInput <- reactive({
    switch(input$dataset,
           "total" = input_dataset,
           "ams" = ams,
           "utr" = utr,
           "ein" = ein)
  })

  # Creation of medians plot
  output$medianPlot <- renderPlot({

    data <- datasetInput()
```

```
  dataset <- data[complete.cases(data),]
  index <- which(duplicated(dataset$ID,fromLast=T))
  v <- 1:nrow(dataset)
  z <- v[-index]
  dataset <- dataset[-z,]

  # Considering subgroups by type of medicine
  MG20.data <- dataset[dataset$Dos.in.mg==20,]
  MG20 <- MG20.data$Avg.daily.use.in.ddd[is.finite(MG20.data$Avg.daily.use.in.
      ddd)]
  MG25.data <- dataset[dataset$Dos.in.mg==25,]
  MG25 <- MG25.data$Avg.daily.use.in.ddd[is.finite(MG25.data$Avg.daily.use.in.
      ddd)]
  MG50.data <- dataset[dataset$Dos.in.mg==50,]
  MG50 <- MG50.data$Avg.daily.use.in.ddd[is.finite(MG50.data$Avg.daily.use.in.
      ddd)]
  MG100.data <- dataset[dataset$Dos.in.mg==100,]
  MG100 <- MG100.data$Avg.daily.use.in.ddd[is.finite(MG100.data$Avg.daily.use.in
      .ddd)]

  # Calculating the medians (statistic of interest because of skewness)
  median20 <- median(MG20)
  median25 <- median(MG25)
  median50 <- median(MG50)
  median100 <- median(MG100)
  median.mixed <- median(dataset$Avg.daily.use.in.ddd)

  library(plotrix)
  library(MASS)

  x <- 1:4
  y <- c(median20,median25,median50,median100)
  plot(x,y,col="red",xaxt="n",xlab='different␣quantities␣of␣medicine␣in␣mg',
       ylab="average␣daily␣use␣in␣ddd",main="␣Median␣values")
  axis(1, at=1:4, labels=c(20,25,50,100))
  points(y,pch=20,col="blue",lwd=3)
  lines(c(1,2),c(median20,median25),lty=2)
  lines(c(2,3),c(median25,median50),lty=2)
  lines(c(3,4),c(median50,median100),lty=2)
  # It doesn't make sense to plot the CI, cause they are really small, as it can
       be
  # expected for the median
  legend(2,1,c(paste("median␣20mg=",round(median20,digits=3)),
               paste("median␣25mg=",round(median25,digits=3)),
               paste("median␣50mg=",round(median50,digits=3)),
               paste("median␣100mg=",round(median100,digits=3))),bty="n")
  abline(h=median.mixed,lty=2,col="red")
  text(2.5,0.25,paste("mixed␣median=",round(median.mixed,digits=3)),col="red")

})

# Creation of histogram of sold sildenafil
output$histogram <- renderPlot({
  data <- datasetInput()

  dataset <- data[complete.cases(data),]
  index <- which(duplicated(dataset$ID,fromLast=T))
  v <- 1:nrow(dataset)
```

```
  z <- v[-index]
  dataset <- dataset[-z,]

  # Considering subgroups by type of medicine
  MG20.data <- dataset[dataset$Dos.in.mg==20,]
  MGelse.data <- dataset[dataset$Dos.in.mg!=20,]

  par(mfrow=c(2,1))
  truehist(MG20.data$ddd,main="sold␣sildenafil␣20␣MG␣in␣ddd",xlab="ddd")
  truehist(MGelse.data$ddd,main="sold␣sildenafil␣25,␣50,␣100␣MG␣in␣ddd",xlab="
      ddd")
})

# Creation of boxplot of sold sildenafil
output$boxplot <- renderPlot({
  data <- datasetInput()

  dataset <- data[complete.cases(data),]
  index <- which(duplicated(dataset$ID,fromLast=T))
  v <- 1:nrow(dataset)
  z <- v[-index]
  dataset <- dataset[-z,]

  # Considering subgroups by type of medicine
  MG20.data <- dataset[dataset$Dos.in.mg==20,]
  MGelse.data <- dataset[dataset$Dos.in.mg!=20,]

  par(mfrow=c(2,1))
  boxplot(MG20.data$ddd,main="sold␣sildenafil␣20␣mg␣in␣ddd",ylab="ddd",outline =
       input$outliers)
  boxplot(MGelse.data$ddd,main="sold␣sildenafil␣25,␣50,␣100␣mg␣in␣ddd",ylab="ddd
      ",outline = input$outliers)
})

# Creation of weekly use plot
output$weekly.use.plot <- renderPlot({
  data <- datasetInput()

  par(mfrow=c(2,1))

  index <- which(data$Dos.in.mg==20)
  MG20.data <- data[index,]
  index <- which(data$Dos.in.mg!=20|is.na(data$Dos.in.mg)==T)
  MGelse.data <- data[index,]

  # 20 mg
  n20 <- length(unique(MG20.data$ID))
  a20 <- rep(MG20.data$Avg.daily.use.in.mg,MG20.data$Periods)
  b20 <- split(MG20.data$City,MG20.data$ID)
  bb20 <- Filter(length,b20)
  c20 <- numeric(length(bb20))
  for(i in 1:length(bb20)){
    c20[i] <- unique(as.character(bb20[[i]]))
  }
  weekly.data20 <- matrix(a20,ncol=days,byrow=T)
  weeks20 <- data.frame(ID=unique(MG20.data$ID),weekly.data20,City=c20)
  n <- ncol(weeks20)
  sums20 <- apply(weeks20[,-c(1,n)],2,sum)
```

```
    N20 <- length(sums20)
    S20 <- numeric()
    j <- 1
    for(i in 1:N20){
      S20[i] <- sum(sums20[j:(j+6)])
      j <- j+7
    }
    ss <- na.omit(S20)
    cumMed <- cummed(ss)
    cumMedinv <- cummed.inverse(ss)
    weeks <- input$weeks
    weeks.inverse <- input$weeks.inv
    plot(ss,type="l",ylab="Sildenafil in mg",xlab="Week",main="sildenafil 20 mg")
    abline(h=cumMed[weeks],col="blue",lty=2)
    abline(h=cumMedinv[weeks.inverse],col="red",lty=2)
    text(20,cumMed[weeks]+2500,paste("weekly median=",round(cumMed[weeks],digits
        =3)),col="blue",font=2)
    text(80,cumMedinv[weeks.inverse]+2500,paste("weekly median=",round(cumMedinv[
        weeks.inverse],digits=3)),col="red",font=2)

    # 25, 50, 100 mg
    n <- length(unique(MGelse.data$ID))
    a <- rep(MGelse.data$Avg.daily.use.in.mg,MGelse.data$Periods)
    b <- split(MGelse.data$City,MGelse.data$ID)
    bb <- Filter(length,b)
    c <- numeric(length(bb))
    for(i in 1:length(bb)){
      c[i] <- unique(as.character(bb[[i]]))
    }
    weekly.data <- matrix(a,ncol=days,byrow=T)
    weeks <- data.frame(ID=unique(MGelse.data$ID),weekly.data,City=c)
    n <- ncol(weeks)
    sums <- apply(weeks[,-c(1,n)],2,sum)

    N <- length(sums)
    S <- numeric()
    k <- 1
    for(i in 1:N){
      S[i] <- sum(sums[k:(k+6)])
      k <- k+7
    }
    ss <- na.omit(S)
    cumMed <- cummed(ss)
    cumMedinv <- cummed.inverse(ss)
    weeks <- input$weeks
    weeks.inverse <- input$weeks.inv
    plot(ss,type="l",ylab="Sildenafil in mg",xlab="Week",main="sildenafil 25, 50
        and 100 mg")
    abline(h=cumMed[weeks],col="blue",lty=2)
    abline(h=cumMedinv[weeks.inverse],col="red",lty=2)
    text(20,cumMed[weeks]+2500,paste("weekly median=",round(cumMed[weeks],digits
        =3)),col="blue",font=2)
    text(80,cumMedinv[weeks.inverse]+2500,paste("weekly median=",round(cumMedinv[
        weeks.inverse],digits=3)),col="red",font=2)
  })

})
```

Script (script.R)

```
# sildenafil App

library(shiny)

input_dataset <- read.csv("sildenafil.csv",header=TRUE,sep=",")
input_dataset$ID <- factor(input_dataset$ID)
input_dataset$Date <- as.Date(input_dataset$Date)
input_dataset$City <- factor(input_dataset$City)
input_dataset$Periods <- as.numeric(input_dataset$Periods)
input_dataset$ddd <- as.numeric(input_dataset$ddd)
input_dataset$Dos.in.mg <- as.numeric(input_dataset$Dos.in.mg)
input_dataset$Avg.daily.use.in.mg <- as.numeric(as.character(input_dataset$Avg.
    daily.use.in.mg))
input_dataset$Avg.daily.use.in.ddd <- as.numeric(as.character(input_dataset$Avg.
    daily.use.in.ddd))

# Data for Amsterdam
ams <- input_dataset[input_dataset$City=="ams",]

# Data for Utrecht
utr <- input_dataset[input_dataset$City=="utr",]

# Data for Eindhoven
ein <- input_dataset[input_dataset$City=="ein",]

first_day <- min(input_dataset$Date)
last_day <- max(input_dataset$Date)
days <- as.numeric(last_day-first_day)

# function for cumulative median
cummed <- function(x){
  n <- length(x)
  y <- rep(0, n)
  for (i in 1:n)
  {
    y[i] = median(x[1:i])
  }
  y
}

cummed.inverse <- function(x){
  n <- length(x)
  y <- rep(0, n)
  for (i in 1:n)
  {
    y[i] = median(x[n:i])
  }
  y
}
```

R code to create the app for methylphenidate.

User Interface (ui.R)

```
# metylphenidate App

shinyUI(fluidPage(

  titlePanel("Use of legal medicines: metylphenidate study"),

  sidebarLayout(
    sidebarPanel(
      selectInput("dataset", "Choose a dataset:",
                  choices = c("total", "ams", "utr", "ein")),

      checkboxInput("outliers", "Show outliers", FALSE),

      sliderInput("weeks", "Choose the ending week (blue line):",
                  min=1, max=80, value=77),

      sliderInput("weeks.inv", "Choose the starting week (red line):",
                  min=1, max=130, value=20),

      img(src="drug-free-america.png",height=300,width=380)
    ),

    mainPanel(
      tabsetPanel(type = "tabs",
                  tabPanel("Plot of median values", plotOutput("medianPlot",height
                       = "1200px")),
                  tabPanel("Histogram of sold metylphenidate", plotOutput("
                      histogram",height = "1200px")),
                  tabPanel("Boxplot of sold metylphenidate", plotOutput("boxplot",
                      height = "800px")),
                  tabPanel("Plot of weekly use", plotOutput("weekly.plot",height =
                       "1200px"))
      )
    )
  )
))
\end{listing}

\textcolor{red}{Server (server.R)}

\begin{lstlisting}
# metylphenidate App

source("script.R")

shinyServer(function(input, output) {

  datasetInput <- reactive({
    switch(input$dataset,
           "total" = input_dataset,
           "ams" = ams,
           "utr" = utr,
           "ein" = ein)
  })

  # Creation of medians plot
```

```
library(plotrix)
library(MASS)

output$medianPlot <- renderPlot({

  data <- datasetInput()

  index <- which(data$Dosage=="cap"|data$Dosage=="tab"|data$Dosage=="tab_mga")
  dataset <- data[-index,]
  par(mfrow=c(3,1))

  # Considering subgroups by type of medicine
  cap.mga.M5.data <- dataset[dataset$Dosage=="METHYLFENIDAAT␣CAPSULE␣MGA␣␣5MG",]
  cap.mga.M5 <- cap.mga.M5.data$Avg.daily.use.in.ddd
  median.cap.mga.M5 <- median(cap.mga.M5)

  cap.mga.M10.data <- dataset[dataset$Dosage=="METHYLFENIDAAT␣CAPSULE␣MGA␣10MG"
      ,]
  cap.mga.M10 <- cap.mga.M10.data$Avg.daily.use.in.ddd
  median.cap.mga.M10 <- median(cap.mga.M10)

  cap.mga.M20.data <- dataset[dataset$Dosage=="METHYLFENIDAAT␣CAPSULE␣MGA␣20MG"
      ,]
  cap.mga.M20 <- cap.mga.M20.data$Avg.daily.use.in.ddd
  median.cap.mga.M20 <- median(cap.mga.M20)

  cap.mga.M30.data <- dataset[dataset$Dosage=="METHYLFENIDAAT␣CAPSULE␣MGA␣30MG"
      ,]
  cap.mga.M30 <- cap.mga.M30.data$Avg.daily.use.in.ddd
  median.cap.mga.M30 <- median(cap.mga.M30)

  cap.mga.M40.data <- dataset[dataset$Dosage=="METHYLFENIDAAT␣CAPSULE␣MGA␣40MG"
      ,]
  cap.mga.M40 <- cap.mga.M40.data$Avg.daily.use.in.ddd
  median.cap.mga.M40 <- median(cap.mga.M40)

  cap.mga <- c(cap.mga.M5,cap.mga.M10,cap.mga.M20,cap.mga.M30,cap.mga.M40)
  median.cap.mga <- median(cap.mga)

  # Plot of the medians
  x <- 1:5
  y <- c(median.cap.mga.M5,median.cap.mga.M10,median.cap.mga.M20,median.cap.mga.
      M30,median.cap.mga.M40)
  plot(x,y,col="red",xaxt="n",xlab='different␣quantities␣of␣methylphenidate␣in␣
      mg',
       ylab="average␣daily␣use␣in␣ddd",main="Capsule␣mga",cex.lab=1.5,cex.main
           =1.5,cex.axis=1.5)
  axis(1, at=1:5, labels=c(5,10,20,30,40),cex.axis=1.5)
  points(y,pch=20,col="blue",lwd=5)
  lines(c(1,2),c(median.cap.mga.M5,median.cap.mga.M10),lty=2)
  lines(c(2,3),c(median.cap.mga.M10,median.cap.mga.M20),lty=2)
  lines(c(3,4),c(median.cap.mga.M20,median.cap.mga.M30),lty=2)
  lines(c(4,5),c(median.cap.mga.M30,median.cap.mga.M40),lty=2)
  legend(1,1.2,c(paste("median␣5mg=",round(median.cap.mga.M5,digits=3)),
                 paste("median␣10mg=",round(median.cap.mga.M10,digits=3)),
                 paste("median␣20mg=",round(median.cap.mga.M20,digits=3)),
                 paste("median␣30mg=",round(median.cap.mga.M30,digits=3)),
                 paste("median␣40mg=",round(median.cap.mga.M40,digits=3))),
```

```
       bty="n",cex=1.5)
abline(h=median.cap.mga,lty=2,col="red")
text(1.5,0.4,paste("median mixed=",round(median.cap.mga,digits=3)),col="red",
    cex=1.5)

# Considering subgroups by type of medicine
tab.M5.data <- dataset[dataset$Dosage=="METHYLFENIDAAT TABLET  5MG",]
tab.M5 <- tab.M5.data$Avg.daily.use.in.ddd
median.tab.M5 <- median(tab.M5)

tab.M10.data <- dataset[dataset$Dosage=="METHYLFENIDAAT TABLET 10MG",]
tab.M10 <- tab.M10.data$Avg.daily.use.in.ddd
median.tab.M10 <- median(tab.M10)

tab.M20.data <- dataset[dataset$Dosage=="METHYLFENIDAAT TABLET 20MG",]
tab.M20 <- tab.M20.data$Avg.daily.use.in.ddd
median.tab.M20 <- median(tab.M20)

tab <- c(tab.M5,tab.M10,tab.M20)
median.tab <- median(tab)

# Plot of the medians
x <- 1:3
y <- c(median.tab.M5,median.tab.M10,median.tab.M20)
plot(x,y,col="red",xaxt="n",xlab='different quantities of methylphenidate in
    mg',
     ylab="average daily use in ddd",main="Tablet",cex.lab=1.5,cex.main=1.5,
         cex.axis=1.5)
axis(1, at=1:3, labels=c(5,10,20),cex.axis=1.5)
points(y,pch=20,col="blue",lwd=5)
lines(c(1,2),c(median.tab.M5,median.tab.M10),lty=2)
lines(c(2,3),c(median.tab.M10,median.tab.M20),lty=2)
legend(2,0.7,c(paste("median 5mg=",round(median.tab.M5,digits=3)),
              paste("median 10mg=",round(median.tab.M10,digits=3)),
              paste("median 20mg=",round(median.tab.M20,digits=3))),bty="n",
       cex=1.5)
abline(h=median.tab,col="red",lty=2)
text(1.2,0.7,paste("median mixed=",round(median.tab,digits=3)),col="red",cex
    =1.5)

# Considering subgroups by type of medicine
tab.mga.M18.data <- dataset[dataset$Dosage=="METHYLFENIDAAT TABLET MGA 18MG",]
tab.mga.M18 <- tab.mga.M18.data$Avg.daily.use.in.ddd
median.tab.mga.M18 <- median(tab.mga.M18)

tab.mga.M27.data <- dataset[dataset$Dosage=="METHYLFENIDAAT TABLET MGA 27MG",]
tab.mga.M27 <- tab.mga.M27.data$Avg.daily.use.in.ddd
median.tab.mga.M27 <- median(tab.mga.M27)

tab.mga.M36.data <- dataset[dataset$Dosage=="METHYLFENIDAAT TABLET MGA 36MG",]
tab.mga.M36 <- tab.mga.M36.data$Avg.daily.use.in.ddd
median.tab.mga.M36 <- median(tab.mga.M36)

tab.mga.M54.data <- dataset[dataset$Dosage=="METHYLFENIDAAT TABLET MGA 54MG",]
tab.mga.M54 <- tab.mga.M54.data$Avg.daily.use.in.ddd
median.tab.mga.M54 <- median(tab.mga.M54)

tab.mga <- c(tab.mga.M18,tab.mga.M27,tab.mga.M36,tab.mga.M54)
median.tab.mga <- median(tab.mga)
```

```
  # Plot of the medians
  x <- 1:4
  y <- c(median.tab.mga.M18,median.tab.mga.M27,median.tab.mga.M36,median.tab.mga
      .M54)
  plot(x,y,col="red",xaxt="n",xlab='different␣quantities␣of␣methylphenidate␣in␣
      mg',
       ylab="average␣daily␣use␣in␣ddd",main="Tablet␣mga",cex.lab=1.5,cex.main
           =1.5,cex.axis=1.5)
  axis(1, at=1:4, labels=c(18,27,36,54),cex.axis=1.5)
  points(y,pch=20,col="blue",lwd=5)
  lines(c(1,2),c(median.tab.mga.M18,median.tab.mga.M27),lty=2)
  lines(c(2,3),c(median.tab.mga.M27,median.tab.mga.M36),lty=2)
  lines(c(3,4),c(median.tab.mga.M36,median.tab.mga.M54),lty=2)
  legend(1,1.85,c(paste("median␣18mg=",round(median.tab.mga.M18,digits=3)),
                  paste("median␣27mg=",round(median.tab.mga.M27,digits=3)),
                  paste("median␣36mg=",round(median.tab.mga.M36,digits=3)),
                  paste("median␣54mg=",round(median.tab.mga.M54,digits=3))),
         bty="n",cex=1.5)
  abline(h=median.tab.mga,lty=2,col="red")
  text(1.55,1.2,paste("median␣mixed=",round(median.tab.mga,digits=3)),col="red",
      cex=1.5)

})

# Creation of histograms of sold metylphenidate
output$histogram <- renderPlot({
  data <- datasetInput()

  index <- which(data$Dosage=="cap"|data$Dosage=="tab"|data$Dosage=="tab_mga")
  dataset <- data[-index,]
  par(mfrow=c(3,1))

  capsule_mga <- dataset[dataset$Dosage=="METHYLFENIDAAT␣CAPSULE␣MGA␣␣5MG"|
      dataset$Dosage=="METHYLFENIDAAT␣CAPSULE␣MGA␣10MG"|dataset$Dosage==""
      METHYLFENIDAAT␣CAPSULE␣MGA␣20MG"|dataset$Dosage=="METHYLFENIDAAT␣CAPSULE␣
      MGA␣30MG"|dataset$Dosage=="METHYLFENIDAAT␣CAPSULE␣MGA␣40MG",]

  truehist(capsule_mga$ddd,xlab="ddd",main="Capsule␣mga",cex.lab=1.5,cex.main
      =1.5,cex.axis=1.5)

  tablet <- dataset[dataset$Dosage=="METHYLFENIDAAT␣TABLET␣␣5MG"|dataset$Dosage
      =="METHYLFENIDAAT␣TABLET␣10MG"|dataset$Dosage=="METHYLFENIDAAT␣TABLET␣20MG"
      ,]

  truehist(tablet$ddd,xlab="ddd",main="Tablet",cex.lab=1.5,cex.main=1.5,cex.axis
      =1.5)

  tablet.mga <- dataset[dataset$Dosage=="METHYLFENIDAAT␣TABLET␣MGA␣18MG"|dataset
      $Dosage=="METHYLFENIDAAT␣TABLET␣MGA␣27MG"|dataset$Dosage=="METHYLFENIDAAT␣
      TABLET␣MGA␣36MG"|dataset$Dosage=="METHYLFENIDAAT␣TABLET␣MGA␣54MG",]

  truehist(tablet.mga$ddd,xlab="ddd",main="Tablet␣mga",cex.lab=1.5,cex.main=1.5,
      cex.axis=1.5)
})

# Creation of boxplots of sold metylphenidate
output$boxplot <- renderPlot({
  data <- datasetInput()
```

```
  index <- which(data$Dosage=="cap"|data$Dosage=="tab"|data$Dosage=="tab_mga")
  dataset <- data[-index,]
  par(mfrow=c(3,1))

  capsule_mga <- dataset[dataset$Dosage=="METHYLFENIDAAT␣CAPSULE␣MGA␣␣5MG"|
      dataset$Dosage=="METHYLFENIDAAT␣CAPSULE␣MGA␣10MG"|dataset$Dosage==""
      METHYLFENIDAAT␣CAPSULE␣MGA␣20MG"|dataset$Dosage=="METHYLFENIDAAT␣CAPSULE␣
      MGA␣30MG"|dataset$Dosage=="METHYLFENIDAAT␣CAPSULE␣MGA␣40MG",]

  boxplot(capsule_mga$ddd,ylab="ddd",main="Capsule␣mga",cex.lab=1.5,cex.main
      =1.5,cex.axis=1.5,outline = input$outliers)

  tablet <- dataset[dataset$Dosage=="METHYLFENIDAAT␣TABLET␣␣5MG"|dataset$Dosage
      =="METHYLFENIDAAT␣TABLET␣10MG"|dataset$Dosage=="METHYLFENIDAAT␣TABLET␣20MG"
      ,]

  boxplot(tablet$ddd,ylab="ddd",main="Tablet",cex.lab=1.5,cex.main=1.5,cex.axis
      =1.5,outline = input$outliers)

  tablet.mga <- dataset[dataset$Dosage=="METHYLFENIDAAT␣TABLET␣MGA␣18MG"|dataset
      $Dosage=="METHYLFENIDAAT␣TABLET␣MGA␣27MG"|dataset$Dosage=="METHYLFENIDAAT␣
      TABLET␣MGA␣36MG"|dataset$Dosage=="METHYLFENIDAAT␣TABLET␣MGA␣54MG",]

  boxplot(tablet.mga$ddd,ylab="ddd",main="Tablet␣mga",cex.lab=1.5,cex.main=1.5,
      cex.axis=1.5,outline = input$outliers)
})

# Creation weekly.plot
output$weekly.plot <- renderPlot({
  dataset <- datasetInput()
  par(mfrow=c(3,1))

  # weekly plot metylphenidate capsule mga
  capsule_mga <- dataset[dataset$Dosage=="METHYLFENIDAAT␣CAPSULE␣MGA␣␣5MG"|
      dataset$Dosage=="METHYLFENIDAAT␣CAPSULE␣MGA␣10MG"|dataset$Dosage==""
      METHYLFENIDAAT␣CAPSULE␣MGA␣20MG"|dataset$Dosage=="METHYLFENIDAAT␣CAPSULE␣
      MGA␣30MG"|dataset$Dosage=="METHYLFENIDAAT␣CAPSULE␣MGA␣40MG"|dataset$Dosage
      =="cap",]
  n <- length(unique(capsule_mga$ID))
  a <- rep(capsule_mga$Avg.daily.use.in.mg,capsule_mga$Periods)
  b <- split(capsule_mga$City,capsule_mga$ID)
  bb <- Filter(length,b)
  c <- numeric(length(bb))
  for(i in 1:length(bb)){
    c[i] <- unique(as.character(bb[[i]]))
  }
  weekly.data <- matrix(a,ncol=days_cap,byrow=T)
  weeks <- data.frame(ID=unique(capsule_mga$ID),weekly.data,City=c)
  n <- ncol(weeks)
  sums <- apply(weeks[,-c(1,n)],2,sum)

  N <- length(sums)
  S <- numeric()
  j <- 1
  for(i in 1:N){
    S[i] <- sum(sums[j:(j+6)])
    j <- j+7
  }
```

```
ss <- na.omit(S)
cumMed <- cummed(ss)
cumMedinv <- cummed.inverse(ss)
weeks <- input$weeks
weeks.inverse <- input$weeks.inv
plot(ss,type="l",ylab="Metylphenidate␣in␣mg",xlab="Week",main="metylphenidate␣
    capsule␣mga",cex.lab=1.5,cex.main=1.5,cex.axis=1.5)
abline(h=cumMed[weeks],col="blue",lty=2)
abline(h=cumMedinv[weeks.inverse],col="red",lty=2)
text(10,cumMed[weeks]+20000,paste("weekly␣median=",round(cumMed[weeks],digits
    =3)),col="blue",cex=1.5,font=2)
text(70,cumMedinv[weeks.inverse]+20000,paste("weekly␣median=",round(cumMedinv[
    weeks.inverse],digits=3)),col="red",font=2,cex=1.5)

# weekly plot metylphenidate tablet
tablet <- dataset[dataset$Dosage=="METHYLFENIDAAT␣TABLET␣␣5MG"|dataset$Dosage
    =="METHYLFENIDAAT␣TABLET␣10MG"|dataset$Dosage=="METHYLFENIDAAT␣TABLET␣20MG"
    |dataset$Dosage=="tab",]
n <- length(unique(tablet$ID))
a <- rep(tablet$Avg.daily.use.in.mg,tablet$Periods)
b <- split(tablet$City,tablet$ID)
bb <- Filter(length,b)
c <- numeric(length(bb))
for(i in 1:length(bb)){
  c[i] <- unique(as.character(bb[[i]]))
}
weekly.data <- matrix(a,ncol=days_tab,byrow=T)
weeks <- data.frame(ID=unique(tablet$ID),weekly.data,City=c)
n <- ncol(weeks)
sums <- apply(weeks[,-c(1,n)],2,sum)

N <- length(sums)
S <- numeric()
j <- 1
for(i in 1:N){
  S[i] <- sum(sums[j:(j+6)])
  j <- j+7
}
ss <- na.omit(S)
cumMed <- cummed(ss)
cumMedinv <- cummed.inverse(ss)
weeks <- input$weeks
weeks.inverse <- input$weeks.inv
plot(ss,type="l",ylab="Metylphenidate␣in␣mg",xlab="Week",main="metylphenidate␣
    tablet",cex.lab=1.5,cex.main=1.5,cex.axis=1.5)
abline(h=cumMed[weeks],col="blue",lty=2)
abline(h=cumMedinv[weeks.inverse],col="red",lty=2)
text(10,cumMed[weeks]+30000,paste("weekly␣median=",round(cumMed[weeks],digits
    =3)),col="blue",cex=1.5,font=2)
text(70,cumMedinv[weeks.inverse]+30000,paste("weekly␣median=",round(cumMedinv[
    weeks.inverse],digits=3)),col="red",font=2,cex=1.5)

# weekly plot metylphenidate tablet mga
tablet_mga <- dataset[dataset$Dosage=="METHYLFENIDAAT␣TABLET␣MGA␣18MG"|dataset
    $Dosage=="METHYLFENIDAAT␣TABLET␣MGA␣27MG"|dataset$Dosage=="METHYLFENIDAAT␣
    TABLET␣MGA␣36MG"|dataset$Dosage=="METHYLFENIDAAT␣TABLET␣MGA␣54MG"|dataset$
    Dosage=="tab_mga",]
n <- length(unique(tablet_mga$ID))
a <- rep(tablet_mga$Avg.daily.use.in.mg,tablet_mga$Periods)
```

```
    b <- split(tablet_mga$City,tablet_mga$ID)
    bb <- Filter(length,b)
    c <- numeric(length(bb))
    for(i in 1:length(bb)){
      c[i] <- unique(as.character(bb[[i]]))
    }
    weekly.data <- matrix(a,ncol=days_tab.mga,byrow=T)
    weeks <- data.frame(ID=unique(tablet_mga$ID),weekly.data,City=c)
    n <- ncol(weeks)
    sums <- apply(weeks[,-c(1,n)],2,sum)

    N <- length(sums)
    S <- numeric()
    j <- 1
    for(i in 1:N){
      S[i] <- sum(sums[j:(j+6)])
      j <- j+7
    }
    ss <- na.omit(S)
    cumMed <- cummed(ss)
    cumMedinv <- cummed.inverse(ss)
    weeks <- input$weeks
    weeks.inverse <- input$weeks.inv
    plot(ss,type="l",ylab="Metylphenidate␣in␣mg",xlab="Week",main="metylphenidate␣
        tablet␣mga",cex.lab=1.5,cex.main=1.5,cex.axis=1.5)
    abline(h=cumMed[weeks],col="blue",lty=2)
    abline(h=cumMedinv[weeks.inverse],col="red",lty=2)
    text(10,cumMed[weeks]+30000,paste("weekly␣median=",round(cumMed[weeks],digits
        =3)),col="blue",cex=1.5,font=2)
    text(70,cumMedinv[weeks.inverse]+30000,paste("weekly␣median=",round(cumMedinv[
        weeks.inverse],digits=3)),col="red",font=2,cex=1.5)
  })

})
```

Script (script.R)

```
# metylphenidate App

library(shiny)

input_dataset <- read.csv2("methylphenidate.csv",header=TRUE,sep=",")
input_dataset$ID <- factor(input_dataset$ID)
input_dataset$Date <- as.Date(input_dataset$Date)
input_dataset$City <- factor(input_dataset$City)
input_dataset$Periods <- as.numeric(input_dataset$Periods)
input_dataset$ddd <- as.numeric(input_dataset$ddd)
input_dataset$Dos.in.mg <- as.numeric(input_dataset$Dos.in.mg)
input_dataset$Avg.daily.use.in.mg <- as.numeric(as.character(input_dataset$Avg.
    daily.use.in.mg))
input_dataset$Avg.daily.use.in.ddd <- as.numeric(as.character(input_dataset$Avg.
    daily.use.in.ddd))
input_dataset$Dosage <- factor(input_dataset$Dosage)

# Distinction per city

# Data for Amsterdam
ams <- input_dataset[input_dataset$City=="ams",]
```

```
# Data for Utrecht
utr <- input_dataset[input_dataset$City=="utr",]

# Data for Eindhoven
ein <- input_dataset[input_dataset$City=="ein",]

capsule_mga <- input_dataset[input_dataset$Dosage=="METHYLFENIDAAT␣CAPSULE␣MGA␣␣5
    MG"|input_dataset$Dosage=="METHYLFENIDAAT␣CAPSULE␣MGA␣10MG"|input_dataset$
    Dosage=="METHYLFENIDAAT␣CAPSULE␣MGA␣20MG"|input_dataset$Dosage=="METHYLFENIDAAT
    ␣CAPSULE␣MGA␣30MG"|input_dataset$Dosage=="METHYLFENIDAAT␣CAPSULE␣MGA␣40MG"|
    input_dataset$Dosage=="cap",]
first_day_cap <- min(capsule_mga$Date)
last_day_cap <- max(capsule_mga$Date)
days_cap <- as.numeric(last_day_cap-first_day_cap)

tablet <- input_dataset[input_dataset$Dosage=="METHYLFENIDAAT␣TABLET␣␣5MG"|input_
    dataset$Dosage=="METHYLFENIDAAT␣TABLET␣10MG"|input_dataset$Dosage=="
    METHYLFENIDAAT␣TABLET␣20MG"|input_dataset$Dosage=="tab",]
first_day_tab <- min(tablet$Date)
last_day_tab <- max(tablet$Date)
days_tab <- as.numeric(last_day_tab-first_day_tab)

tablet_mga <- input_dataset[input_dataset$Dosage=="METHYLFENIDAAT␣TABLET␣MGA␣18MG"
    |input_dataset$Dosage=="METHYLFENIDAAT␣TABLET␣MGA␣27MG"|input_dataset$Dosage=="
    METHYLFENIDAAT␣TABLET␣MGA␣36MG"|input_dataset$Dosage=="METHYLFENIDAAT␣TABLET␣
    MGA␣54MG"|input_dataset$Dosage=="tab_mga",]
first_day_tab.mga <- min(tablet_mga$Date)
last_day_tab.mga <- max(tablet_mga$Date)
days_tab.mga <- as.numeric(last_day_tab.mga-first_day_tab.mga)

# function for cumulative median
cummed <- function(x){
  n <- length(x)
  y <- rep(0, n)
  for (i in 1:n)
  {
    y[i] = median(x[1:i])
  }
  y
}

cummed.inverse <- function(x){
  n <- length(x)
  y <- rep(0, n)
  for (i in 1:n)
  {
    y[i] = median(x[n:i])
  }
  y
}
```