# CHRIS Study

# Drug information

Version 1.1

24th April 2024

Authors: LB, LF, MG

## 1. Introduction

This module stores the drugs of the participants, whose packages were scanned at the CHRIS study center.

Participants book a morning appointment at the CHRIS study center, ranging from 7.45 to 8.45 a.m. Each study participant is assigned a workflow at the reception. If there are ten study participants (maximum capacity), there are ten different workflows, marked with the letters from “A“ to “K“. The current workflow is as follows: A-B-C-D-E-F-G-H-I-K. All the workflows can be found in the documentation of CHRIS Baseline/General information/Administrative data, in the file named “Workflows at baseline assessment”. The drugs are scanned after the neurological tests, which occur after the interview.

Participants are asked to bring all the packages of the medications they have taken in the last 7 days or that they take regularly.

For each drug package that was brought by the participant, the interviewer had to ask a set of questions on the dosage, prescription, route of administration, and the therapy length. More precisely, the participant is first asked “*In the last 7 days have you taken drugs, or food integrators as vitamins or minerals? Please also consider anti-inflammatory and painkiller drugs, insulin preparations, galenics and injections. Please indicate drugs with prolonged action, even if they have been taken more than 7 days ago. Also consider preparations bought at the supermarket/grocery*.”

If drug usage is reported, then other information on drugs are entered in the database.

Among all the information collected, the relevant variable to classify drugs is the ATC (Anatomical Therapeutic Chemical Classification System) code, which has been extensively cleaned and assigned to all the drugs.

In the ATC classification, the active substances are divided into different groups according to the organ or system on which they act and their therapeutic, pharmacological, and chemical properties.

There are, however, cases when the drug information is inserted by hand, specifically when

- the drug package is scanned with the barcode reader, but it is not found (database not updated, or drug from abroad),
- the drug package is forgotten by the participant, and just the name is reported.

In these cases, operators can enter free text and choose among drugs in a scrolling list.
Therefore, some codes with AIC (“*Autorizzazione all’Immissione al Commercio*”) and an ATC seem apparently scanned but were actually inserted by hand. It is not possible to know which records fulfill this condition. This type of data entry might be responsible of additional heterogeneity observed (e.g. in pharmacological forms in R03 drugs). If a record is not found in the scrolling list, it is inserted by hand, and the quality of information is highly variable.

Therefore, the drug database has one or more lines per participant, depending on the number of drugs taken. If no drug usage was reported, then there is a single line for this participant. If instead drug usage was reported, then for this participant there are as many lines as the number of reported drugs.

The guidelines for scanning the drugs package are available at CHRIS Baseline/Drugs data in the documents named “Drugs data insertion guidelines”.

## 2. History version changes

The drug data were collected since August 24th, 2011 and no version change occurred. The cleaning process has produced the following changes:

**variables added:** x0dd02a, x0dd04, x0dd05, x0dd06, x0dd09, x0dd10a, x0dd16, x0dd18a

## 3. Data cleaning

1. The CHRIS drug dataset was loaded.
2. The drug database was merged with age and sex information of the participants, using the participation number aid as key variable, to enhance later plausibility check.
3. Duplicate of rows with the same participant number aid and the same time of insertion were dropped.
4. The consistency of the scanned ATC code and the drug usage x0dd01 was checked. If the drug usage was “No”, the ATC code had to be empty. Aberrant values of ATC code, e.g. not starting with a letter and therefore correspondent to integrators and homeopathic treatments, were checked. A copy of the scanned ATC code was created only if the first two characters were a letter followed by a number. It was saved as x0dd03.
5. A variable classifying the active compound of the drug was derived, looking for every ATC code and commercial drug name. It was saved as x0dd06. For each therapeutic subgroup (first 3 characters of the ATC code x0dd03), each active compound was reassigned the correct ATC code.
6. Active compounds with different anatomical applications or different pharmacological forms were assigned separately the ATC code.
7. When the drug name contained multiple active compounds or commercial drug names, it was either assigned a single active compound and ATC code if a medicament containing all the mentioned active compounds exist or expanded into as many rows and different ATC code as the active compounds mentioned in x0dd03 (e.g., “anticoagulants, beta blocker, antihypertensive, cholesterol drug” was split into 4 rows with active compounds, respectively, “Antithrombotic agents”, “Beta blocking agents”, “Cardiovascular system”, and “Lipid modifying agents”).
8. For drug names of very low quality, the ATC code and the active compound were recovered if the name allowed it, at least partially (e.g., “Asthma spray” had as active compound “Drugs for obstructive airway diseases” and ATC “R03”). If the name did not allow it, e.g. just mentioning the drug producer and a dosage, then both the active compound variable x0dd06 and the ATC code x0dd03 were set to “Unexpected missing” (-89).
9. Additionally quality checks to verify the proper assignment of ATC codes included:
    a) drugs with available cleaned ATC code (x0dd03) could not have the active compound name x0dd06 missing,
    b) if the active compound x0dd06 was missing, original ATC code had to be missing or starting with a number,

   c) drugs with neither scanned ATC, nor cleaned ATC code in x0dd03, that were probably inserted manually and had typos in the drug commercial name x0dd07, were saved separately to be corrected manually.

10. The drugs whose scanned ATC code is different from the cleaned ATC code, and available active compound (so nor integrators nor homeopathic treatments) were assessed to identify possibly wrong codes and to count how many ATC codes were corrected.
11. The drugs in the ATC class G (Genito urinary system and sex hormones) were checked if they were consistent with age and sex of the participants. If the drug was not consistent with the gender, its active compound and ATC code were set to “Unexpected missing”.
12. The main ATC class (the first character) was extracted from the cleaned ATC code x0dd03 and saved as x0dd04.
13. The ATC therapeutic subgroup (the first two characters) was extracted from the cleaned ATC code x0dd03 and saved as x0dd05.
14. Special cases of drug usage were assessed looking at the comments of the study assistants in x0dd18. Special cases include:
    a) The only drug used were anesthetics at the dentist or at the hospital, in this case ATC code and active compound were set to “Unexpected missing” (-89),
    b) The only drug reported were administered at the emergency hospital, also in this case ATC code and active compound were set to “Unexpected missing”,
    c) The only drug reported is homeopathic, then the ATC code and active compound were set to “Missing by design” (-99),
    d) The only drug reported was a contrast medium, also in this case ATC code and active compound were set to “Unexpected missing”.
15. A variable on drug usage, with a single value per participant, was created to consider the participants regularly taking actual drugs as “Yes”, and all the special cases mentioned above as lack of drug usage, i.e. with value “No”. This variable was saved as x0dd02a.
16. A variable on the data cleaning was created to classify the special cases mentioned at point 14 and saved as x0dd16.
17. A copy of the scanned pharmacological form was created, with the following values:
    a) “Yes” if the participant regularly used drugs (x0dd02a=”Yes”),
    b) “Missing by design” (-99) if the participant did not use any drug (x0dd02a=”No”),
    c) “Unexpected missing” (-89) if the information on drug usage was missing (x0dd02a=-89).

    It was saved as x0dd08.

18. A new variable, x0dd09, was created to classify the route of administration of each drug:
    a) “Missing by design” (-99) if the pharmacological form is “Missing by design” (x0dd08=-99),
    b) "Unexpected missing” (-89) if the pharmacological form is “Unexpected missing” (x0dd08=-89),
    c) For each therapeutic subgroup and feasible pharmacological form, a route of administration was assigned, between the categories "oral”, "respiratory”, "parenteral", "conjunctival", "topical", "rectal", "intravesical", "oropharyngeal", "transdermal", "vaginal", "intrauterine”, "sublingual", "auricular", "nasal", "ophthalmic", "subcutaneous", "infiltration", "intravenous", "intrathecal", "intra-articular", and

"intragingival". Reference database for the types of route of administration is https://www.fda.gov/drugs/data-standards-manual-monographs/route-administration

19. For some drugs with a route of administration missing in x0dd09 but an available ATC code in x0dd03, the route of administration was recovered according to the ATC treatment subgroup:
    a) “oral” was assigned if the main class of ATC was “A - Alimentary tract and metabolism” or “N - Nervous system” or “M Musculo-skeletal system”, or if the treatment subgroup was “Cough and cold preparations” (R05), also depending on the commercial drug name
    b) “respiratory” was assigned if the treatment subgroup was “Drugs for obstructive airway diseases” (R03),
    c) “intragingival” if the pharmacological form was “injection” occurred at the dentist (from text in x0dd18),
    d) “Unexpected missing” (-89) if the drug name x0dd07 was a contrast media.
20. The variable x0dd02a, classifying the type of each drug was modified with these values:
    a) “At least one ATC drug” if the participant reported at least one drug with ATC starting with a letter,
    b) “Food integrators only” if the participant reported only food integrators,
    c) “Homeopathy only” if the participant reported only homeopathic products, that were identified by ATC starting with 2, the homeopathic-only drug producers, and drug names,
    d) “Mix of food integrators and homeopathy” if the participant reported both food integrators and homeopathic products.
21. The taking mode variable x0dd12 on regular use of the drug was assigned the values:
    a) “Unexpected missing” (-89) if the drug type was missing (x0dd02a=-89) or the taking mode was missing despite drug usage (x0dd12=. and x0dd02a=”At least one ATC drug”),
    b) “Missing by design” (-99) if there was no drug usage (x0dd02a!=”At least one ATC drug” and x0dd02a!=-89),
    c) The reported taking mode was kept in any other case.

    In the data cleaning procedure, priority was given to drugs with an ATC code, leaving uncleaned food integrators and homeopathy. In most cases, records are low quality and it is very difficult to establish the nature of the product (e.g. Magnesium).

22. The taking interval variable x0dd13 was assigned the values:
    a) “Unexpected missing” (-89) if the drug type was missing (x0dd02a=-89) or the taking interval was missing despite drug usage (x0dd13=. and x0dd02a=”At least one ATC drug”),
    b) “Missing by design” (-99) if there was no drug usage (x0dd02a!=”At least one ATC drug” and x0dd02a!=-89),
    c) The reported taking interval was kept in any other case.
23. The taking period variable x0dd14 was assigned the values:
    a) “Unexpected missing” (-89) if the drug type was missing (x0dd02a=-89) or the taking period was missing despite drug usage (x0dd14=. and x0dd02a=”At least one ATC drug”),
    b) “Missing by design” (-99) if there was no drug usage (x0dd02a!=”At least one ATC drug” and x0dd02a!=-89),

c) The reported taking period was kept in any other case.

24. The taken today variable x0dd15 was assigned the values:

a) “Yes” if the participant reported to have taken it that day or the pharmacological form was IUD, vaginal ring or implant,

b) “No” if the participant did not report to have taken it that day and the pharmacological form was not IUD, vaginal ring nor implant,

c) “Unexpected missing” (-89) if the drug type was missing (x0dd02a=-89) or the taken today variable was missing despite drug usage (x0dd15=. and x0dd02a=”At least one ATC drug”),

d) “Missing by design” (-99) if there was no drug usage (x0dd02a!=”At least one ATC drug” and x0dd02a!=-89).

25. The decreed variable x0dd11 was assigned the values:

a) “Unexpected missing” (-89) if the drug type was missing (x0dd02a=-89) or the decreed variable was missing despite drug usage (x0dd11=. and x0dd02a=”At least one ATC drug”),

b) “Missing by design” (-99) if there was no drug usage (x0dd02a!=”At least one ATC drug” and x0dd02a!=-89),

c) The reported information on prescription was kept in any other case.

26. The variables related to the ATC code, x0dd03-x0dd05, had their values changed in the following cases:

a) They were set to “Unexpected missing” (-89) if the drug type x0dd02a was missing (-89) or if the participant used drugs but had empty information on cleaned ATC and its active compound (x0dd02a=”At least on ATC drug” and x0dd06=”-89” and x0dd03=”-89”),

b) They were set to “Missing by design” (-99) if the drug type x0dd02a was neither missing nor if the participant used drugs (x0dd02a!=”At least on ATC drug” and x0dd02a!=-89).

The same procedure was followed for pharmacological form and rout of administration if they were still missing.

27. The active compound variable x0dd06 has its missing observations set to:

a) “Unexpected missing” (-89) if the cleaned ATC was “Unexpected missing” (x0dd03=”-89”),

b) “Missing by design” (-99) if the cleaned ATC was missing by design (x0dd03=”-99”),

28. The dosage variable, x0dd10, has been translated and categorized, when possible, into the variable x0dd10a, taking into account the correct pharmacological form stored in x0dd08. It still does not contain the amount of active compound of the daily dosage, but rather the quantities of the pharmacological form the participants regularly take. It was not always possible to determine the unit the participant was referring to.

29. The variables storing the notes with the comments of the study assistants, x0dd18, was translated and categorized, when possible, into the variable x0dd18a.

30. The CHRIS drug dataset was saved.

## 4. Advices for the analysis

The content of the nurse's notes includes information on drug acquired abroad, dosages that are not constant (e.g. thyroid therapies), treatments that were discontinued (for a pregnancy or a drug change), and dosages split in multiple times of the day.

The **x0dd01** binary indicator is "yes" if the participant declares some drug intake in the week before the interview. However, the actual number of participants taking at least one **drug with ATC code** is actually lower than the number of those answering "yes". Furthermore, 4% of participants answer "yes", but did not report any other information. The new drug indicator **x0dd02a** shows the actual drug intake per participant, excluding the reports of only integrators or homeopathic treatments.

The variable x0dd11 (Drug prescribed by the doctor) looks unreliable. In fact, among "recommended" drugs *or* drugs without prescriptions, there are drugs which typically *need* a prescription. Because it is not possible to distinguish among putative data entry mistakes, or real cases where the drug might be bought without prescription, x0dd11 is almost unchanged and should not be used.

The drug name variable, x0dd07, represents the commercial name of the drug and for the same active compound, x0dd06, it is possible to find many different drug names. The commercial drug name has been used to check ATC codes and to derive the active compound variable (plain or in combination with other substances).

**Drug dosage** (x0dd10) is a complex variable due to the following reasons:

- Dosage includes mixed unit of measures and mixed statistical units. Dosage = "1" could either correspond to 1 tablet or 1 g or 1 mg of active compound. Open text is also present.
- To determine the dosage of drugs in combination, each record should be expanded, so that each active compound is described by a line. For example, **OLPREZIDE 28CPR RIV 20MG+12,5M** (ATC=C09DA08) is a combi drug containing Olmesartan medoxomil and diuretics. To determine the dosage of each active compound the record should be split in **"C09CA08, Olmesartan medoxomil"** and **"C03AA03, Hydrochlorothiazide"**. Therefore, the dataset must be reshaped.
- For drugs requiring therapeutic plans, e.g. warfarin, the reported dosage is valid for a limited time (e.g. the ongoing week). In case of warfarin for example, therapy is continuously adjusted depending on blood coagulation values. This intrinsic data limitation must be recognized.
- For pharmacological forms different from tablets, capsules, injections, or any other pharmacological form delivering a fixed amount of compound at a fixed time, it is difficult to establish the exact amount of drug intake. Daily dosage cannot be established for creams, ointments, drops, powders for inhalation (e.g. in case of asthma, used when needed), implants / IUD (continuously releasing drugs over time by definition). Daily dosage can also not always be established for some insulins (used when needed).

In the variable x0dd10a the correct pharmacological form x0dd08 was added at the end of the x0dd10 text when it was clear what the unit was meant (e.g. tablet or capsules would be with integers or fractions, whereas ointments, sprays, and drops were more difficult to assign).

Furthermore, in rare occasions the information reported in the interview might disagree with the drugs scanned (or lack thereof) at the CHRIS center: either some participants reported chronic diseases that are regularly treated, but did not bring the relative drug package, or other participants might have omitted to suffer from severe diseases at the interview and have brought drug packages that are highly specific of a severe disease.

Finally, the analyst should always take into account that the operator in charge of scanning the medication package and asking the relative questions might have influenced how the participant reported their answers. The analyst should therefore adjust for the study assistant variable, x0dd19, when possible.

## 5. References

Information about ATC codes can be found at http://www.whocc.no/atc/structure_and_principles/

ATCs can be found at http://www.whocc.no/atc_ddd_index/

Route of administration database: https://www.fda.gov/drugs/data-standards-manual-monographs/route-administration