**S5 File. Expanded methods**

**1. The R programme to obtain data from CTRI, and store it in an SQLite database.**

The CTRI database hosts records of many trials. Each record is in the form of a table. In order to analyze the data pertaining to each relevant record in CTRI, we wrote an R program to download and scrape the data. However, there are variabilities in the tabular format. Illustratively, subtables may be missing in a given record. Such variations created challenges for automating data curation.

Each webpage is identified by the trial CTRI number. The URL of the webpage also contains trial ID, although this ID is not present in the body of the record. The R program looked for the trial ID in the URL of the trial record, and then downloaded the relevant html files. Next, XML package in R read each html file, and presented the data as a dataframe, with 3–4 columns. The raw data was mined to obtain the data relevant to particular fields such as CTRI number, study details, enrollment details, recruitment details, etc. R packages such as readr, stringr and tidyr were used to process the data in these fields. Then, an SQLite database was created inhouse using the RSQLite package, and the processed data stored in this database.

**2. Inconsistencies**

As mentioned above, a given trial record may have had some deviations from the template. We needed to perform string manipulations so that all the data was available in a standard format. Illustratively, the address of a study site should have had the place, the city, the state and the pincode in different rows, but this was not always the case. We needed to properly segment the data to ensure the correct format of each address.

**3. Accessing the data from the database and carrying out data related tasks**

The SQLite database, developed inhouse, had tables dedicated for each field. Each table included the relevant CTRI number as well. As such, it was easy to work with the database, and we used python, R etc. and GUIs like DB Browser for SQLite (DB4S) to access the data. Basic SQL queries were also used, especially to export data from the database. The formulae, functions and pivot tables of Microsoft Excel and LibreOffice Calc were used for further processing of the data.

**4. Examples**

A few examples of the basic queries commonly used with the database are as follows:

| | Description | R Code |
|---|---|---|
| 1. | list all data for recent five year period | trials_btwn_15may16_14may21= subset(common_dataset, registered_on >= "2016-05-15" & registered_on <= "2021-05-14", select = trial_id:reg_type) |
| 2. | list all data for interventional studies | Intrvntl = subset(trials_btwn_15may16_14may21, type_of_trial == " Interventional", select = trial_id:reg_type) |
| 3. | list all data for phase 2 information of the trials | Intrvntl_ph2 = subset(Intrvntl, (phase == " Phase 2")\|(phase ==" Phase 2/ Phase 3"), select = trial_id:reg_type) |
| 4. | list all data for phase 3 information of the trials | Intrvntl_ph3 = subset(Intrvntl, (phase == " Phase 3")\|(phase == " Phase 2/ Phase 3")\|(phase == " Phase 3/ Phase 4")\|(phase ==  " Phase 3, Phase III"), select = trial_id:reg_type) |
| 5. | list all data for phase 3 information of the trials | Intrvntl_ph3 = subset(Intrvntl, (phase == " Phase 3")\|(phase == " Phase 2/ Phase 3")\|(phase == " Phase 3/ Phase 4")\|(phase ==  " Phase 3, Phase III"), select = trial_id:reg_type) |

**5. Removing a duplicate CTRI number**

In S7 File, the total number of trials was reduced from 5454 to 5453, since the same CTRI number (CTRI/2021/04/033023) was allotted to two trials that had different trial IDs, 54077 and 54260. We note that subsequently, trial 54260 was given another CTRI ID (CTRI/2021/04/033024).

**6. How sponsors were named and classified**

Occasionally the names of sponsors or their categories had potential to cause confusion or complicate the analysis. These issues are listed below, along with the steps taken to avoid confusion.

1. Since a given sponsor name may have had variations, a 'standardized name' was arrived at for each sponsor of more than one trial. These are listed in S8 File.

2. There were obvious typos in some sponsor names or classifications. We chose the name that was the closest to the correct name, as the standardized name for that sponsor.

3. If there was a corporate office and R&D or manufacturing unit in the same city, we chose the name that was provided the most often, as the standardized name for that sponsor.

4. Where unrelated institutions with similar names were sponsors, then further identifiers such as the city name were included along with the sponsor name.

5. If there was a chain of hospitals that sponsored multiple trials, we listed each hospital separately. Also, we included the city with the sponsor name (where it was provided).

6. The same sponsor name may have been referred to with its full name or with any of several abbreviations, and sometimes as acronyms. The abbreviations had to be decoded in order to correctly club all the trials sponsored by a given sponsor. When the name was presented best in the address, we took that version of the name.

7. Occasionally, as apparent across trial records, an organization underwent a name change. In cases where this could be confirmed, the newer name was considered.

8. A given institution may have used various names for the sponsor, including the names of departments and of constituent colleges, for instance. In such cases, we took a pair of names as the standardized names, that represent the higher levels of the organization. The list of such pairs of names is provided here:

    a) MAHE / Manipal University
    b) Amrita University / AMRITA VISHWA VIDYAPEETAM
    c) Aligarh Muslim University / Jawaharlal Nehru Medical College
    d) Datta Meghe Institute of Medical sciences / Mahatma Gandhi Ayurved College Hospital and Research Centre
    e) University College of Ayurveda / Dr Sarvepalli Radhakrishnan Rajasthan Ayurved University
    f) Parul Institute Of Ayurved / Parul University
    g) Pt. Bhagwat Dayal Sharma University of Health Sciences, Rohtak / Pt B D Sharma Post Graduate Institute of Medical Sciences
    h) SRM University / SRM Institute of Science and Technology
    i) Shri Krishna Ayush University / SHRI KRISHNA GOVERNMENT AYURVEDIC COLLEGE AND HOSPITAL KURUKSHETRA
    j) Shri Guru Ram Rai University / Shri Guru Ram Rai Institute of Medical and Health Sciences
    k) Sri Dharmasthala Manjunatheshwara College of Ayurveda and Hospital / SDM Ayurveda Hospital
    l) Surya Children Hospital / Surya Childrens Medicare Private limited

9. For each sheet representing sponsor names that began with a given letter, the top of the sheet contains 'SINGLES', that is trials that had unique sponsors. In the initial categorization of trials by sponsor, sometimes similar sponsor names, that were clubbed, later turned out to be different organizations. These were segregated,

but even if they became singles after segregation, they were listed next to the latter set of trials, and not with the SINGLES. (This issue did not affect the analysis in any way.)

10. A company incorporated both in India and abroad was classified as two companies.

11. If the grant name or the 'Institutional Review Board', for instance, was listed as the sponsor, we changed that to the institution's name.

12. In cases where a specific initiative of an organization was listed as the sponsor, we considered the organization as the sponsor.

13. In one case the State government was listed as the sponsor, but actually it appeared to be an institute that was funded by the State government. We considered the institute as the sponsor. The case was as follows: CTRI/2020/01/023081; Government of West Bengal, Institute of Post Graduate Medical Education & Research (IPGMER), Kolkata

14. One sponsor name, the Ministry of Health and Family Welfare, referred to the Government of India (CTRI/2017/06/008731) and to the Tamil Nadu state government (CTRI/2021/04/032773), in two trials. These sponsors were not clubbed under a single sponsor, but were considered separately.

15. Some hospitals were affiliated with, and attached to, other institutes. We made such linkages whenever possible, but may have missed some.