descriptive-analysis

Introduction

This document presents the descriptive results from the analysis of 4 different tools (`sciscore`, `trialidentifier`, `ctregistries` and `nct`). The aim of these tools is to screen papers for one or more registration IDs (protocol, trial registry, etc).

Overall, 200 IDs found in 117 papers were analyzed.

Main analyses:

1. Types of ID detected
2. Where were the IDs located (abstract section, methods section, etc.)?
3. Were the papers research articles?
4. Did the tools agree/disagree? (i.e. they all found the ID = agree).
5. Did the tools find an actual ID?
6. Additional observations

1) Types of ID detected

```
#load dataset
data <- read_csv(here("testing", "regset.csv"))

#general overview
table(data$id_type_group)
```

```
false_positive       protocol            trn
            29              2            169
```

The three main groups of ID observed were `trial registration numbers (trn)`, `protocols`, and `false positives`.

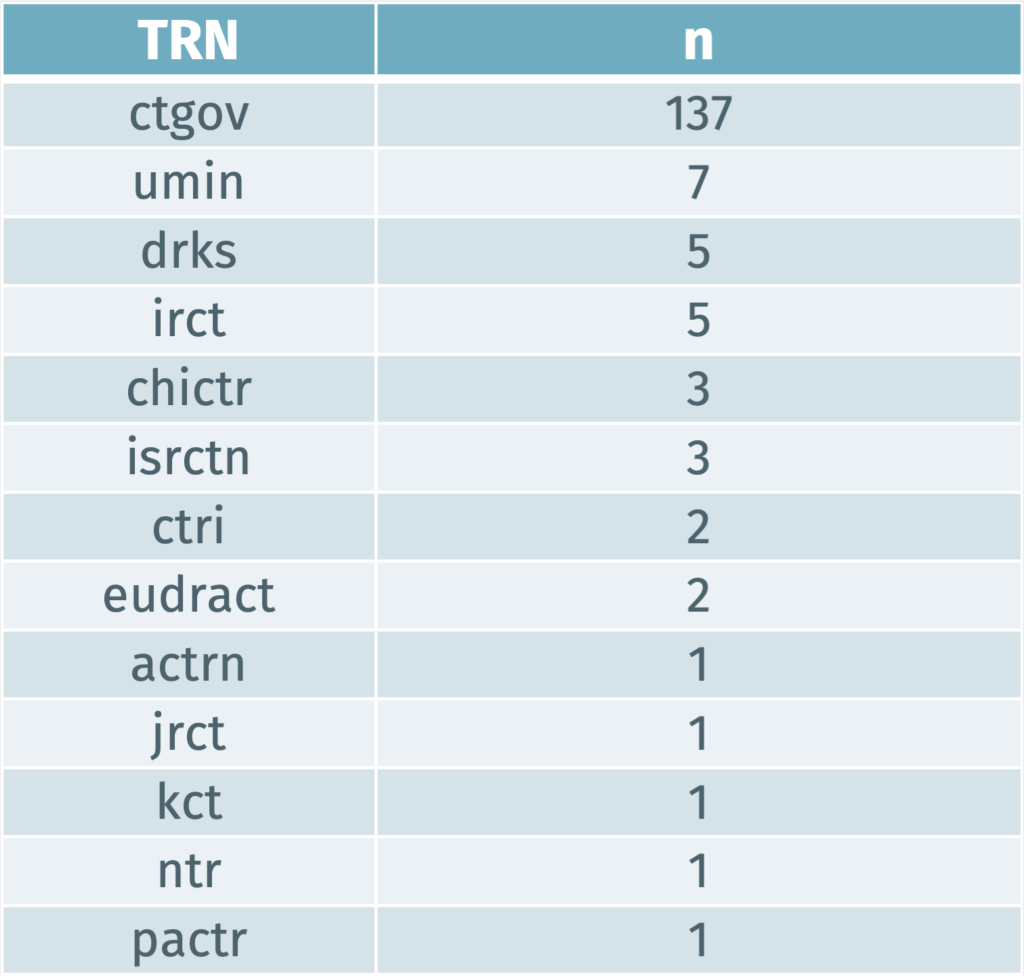
TRNs

```
trn <- subset(data, data$id_type %in% c("ctgov", "umin", "drks", "irct", "chictr",
                                        "isrctn","ctri", "eudract", "actrn", "jrct",
                                        "kct", "ntr", "pactr"))

table(trn$id_type)
```

```
  actrn  chictr   ctgov    ctri    drks eudract    irct  isrctn    jrct     kct 
      1       3     137       2       5       2       5       3       1       1 
    ntr   pactr    umin 
      1       1       7 
```

The majority of the trial registry numbers were from the ctgov registry, followed by umin, dkrs and irct.



TRN	n
ctgov	137
umin	7
drks	5
irct	5
chictr	3
isrctn	3
ctri	2
eudract	2
actrn	1
jrct	1
kct	1
ntr	1
pactr	1

note on the names:

- ctgov - US Clinical Trials Register
- umin - University Hospital Medical Information Network Register
- drks - German Clinical Trials Register
- irct - Iranian Registry of Cinical Trials
- chictr - Chinese Clinical Trial Registry
- isrctn - ISRCTN Registry
- ctri - Clinical Trials Registry-India
- eudract - EU Clinical Trials Register

- actrn - Australian New Zealand Clinical Trial Registry
- jrct - Japan Registry of Clinical Trials
- kct - Korean Clinical Trial Registry
- ntr - Dutch Trial Register
- pactr - Pan African Clinical Trials Registry

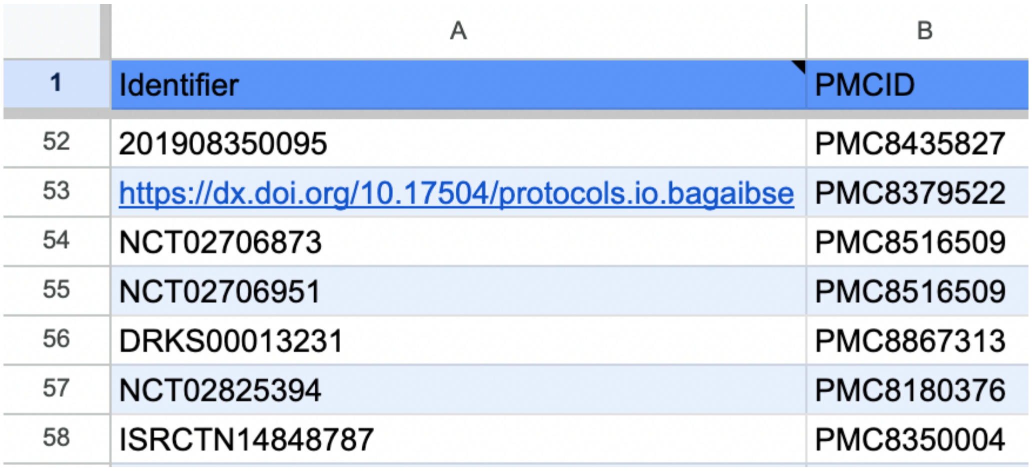
Protocols

```
protocol <- subset(data, data$id_type %in% c("protocols_io"))

table(protocol$id_type)
```

```
protocols_io
           2
```

We only had two IDs that were included in links to specific protocols from protocols.io



	A	B
1	Identifier	PMCID
52	201908350095	PMC8435827
53	https://dx.doi.org/10.17504/protocols.io.bagaibse	PMC8379522
54	NCT02706873	PMC8516509
55	NCT02706951	PMC8516509
56	DRKS00013231	PMC8867313
57	NCT02825394	PMC8180376
58	ISRCTN14848787	PMC8350004

False positives

```
false_positive <- subset(data, data$id_type %in% c("catalogue_id", "datapoint",
                                                    "drug_id", "funding_id", "medical_acron
                                                    "medical_device"))
```

```
table(false_positive$id_type)
```

```
   catalogue_id       datapoint         drug_id      funding_id medical_acronym
              3               5               6              13               1
 medical_device
              1
```

The main types of false positives we encounter were IDs corresponding to specific fundings or grants.

False positive	n
funding_id	13
drug_id	6
datapoint	5
catalogue_id	3
medical_acronym	1
medical_device	1
Total	**29**

2) Where were the IDs located?

```
#id in abstract
table(data$id_in_abstract)
```

```
FALSE  TRUE
  157    43
```

```
#id in methods
table(data$id_in_methods)
```

```
FALSE  TRUE
  106    94
```

```
#id in other location
table(data$id_in_other_location)
```

```
FALSE  TRUE 
   87   113 
```

Location	count
id_in_abstract	43
id_in_methods	94
id_in_other_location	113

What were the other locations? - introduction, discussion, ethic statement, acknowledgements, etc.

```
table(data$other_location)
```

```
              acknowledgements    acknowledgements, footnotes 
                            13                              1 
                  declarations                     disclosure 
                             1                              1 
                    discussion               ethics_statement 
                            50                              8 
                     footnotes                   introduction 
                             1                             18 
introduction, acknowledgements       introduction, discussion 
                             2                              3 
      introduction, trial_info                        results 
                             1                             11 
                    trial_info 
                             2 
```

Note: the IDs analyzed could be in one or more locations, that is why there are more cases (250) out of a total of 200 IDs.

3) Were the papers research articles?

```
table(data$paper_is_research_article)
```

```
FALSE  TRUE
   41   159
```

Here we evaluated if the papers had the structure of a research article (abstract, introduction, methods, results, discussion, etc.). Overall, the majority of the papers were research articles.

It is important to note that 33 of the False cases came from the same paper, which reviewed different studies and identified them using their ctgov trial registry number.

4) Did the tools agree/disagree?

```
table(data$tools_agree)
```

```
 no yes
146  54
```

Overall, the tools agreed in 54 cases, and disagreed in 146. Additionally, they tended to only agree when the ID was a trn:

```
table(data$tools_agree, data$id_type_group)
```

```
      false_positive protocol trn
  no              29        2 115
  yes              0        0  54
```

5) Did the tools find an actual ID?

```r
#subset of sciscore_hit = TRUE
sciscore <- subset(data, data$sciscore_hit %in% c("TRUE"))

table(sciscore$id_type_group)
```

```
protocol      trn
       2       55
```

`sciscore` detected 55 trial registry number IDs and 2 protocol IDs.

```r
#subset of trialidentifier_hit = TRUE
trialidentifier <- subset(data, data$trialidentifier_hit %in% c("TRUE"))

table(trialidentifier$id_type_group)
```

```
trn
139
```

`trialidentifier` identified 139 trial registry number IDs

```r
#subset of ctregistries_hit = TRUE
ctregistries <- subset(data, data$ctregistries_hit %in% c("TRUE"))

table(ctregistries$id_type_group)
```

```
false_positive            trn
            29            166
```

Out of the 4 tools, `ctregistries` was the one who selected the 29 cases of false positives. However, it also detected 166 trial registry number IDs.

```r
#subset of nct_hit = TRUE
nct <- subset(data, data$nct_hit %in% c("TRUE"))
```

```
table(nct$id_type_group)
```

```
trn
137
```

`nct` identified 137 trial registry number IDs.

Overall, it is important to further analyze if the trn IDs identified by `ctregistries` were or not picked up by the other three tools. Even if it picked up false positives, it could also mean that it is less specific and therefore picks other trial registry numbers that are not detected by the other tools.

6) Additional observations

TRN from the study itself or as a reference to other studies

Some papers analyzed use specifically the ctgov trial registry number as a way to cite other studies in their introduction or discussion. Around 71 of the ctgov trn's were used as references. In this case, the tools correctly identify a trial ID number, but is this what we want to find? or are we interested in finding if authors put in their papers their own registry ID?

Funding IDs from China

Sometimes they were detected as IDs, but not in all cases. What were the differences?

TRN with missing digits

There were two cases of ctgov IDs reported with one digit less, but the tools were still able to identify them.