UNIVERSITY COLLEGE LONDON

A Short Demo for the Anonymization Procedure

Sinan Shi

October 26, 2016

The identifiable data set

The identifiable data set is usually stored in ccRecord format. In the following code, we create the ccRecord object from a XML file which contains only five episodes.

```
library(ccdata)
library(ccanonym)

ccd <- xml2Data("../tests/data/test_data_anonym.xml")
demg.table <- as.data.frame(sql.demographic.table(ccd))
```

The YAML configuration file

Create a YAML configuration file as such, where the identifiable variables (dirctVars), key categorical variables (keyVars), key numerical variables (numVars), key date-time variables (datetimeVars), sensitive variables (sensVars) and the corresponding operations and thresholds are specified.

```
directVars:
    - pasno     # PAS number
    - ICNNO     # Site code
    - ADNO      # INCNARC admission number
    - NHSNO     # NHS number
    - TUADNO    # Transferrring unit admission number
    - DOB           # Date of birth

keyVars:
    - GPCODE    # GP code
    - SEX           # Sex
    - PCODE     # Postcode

sensVars:
    - BPC       # Biopsy proven cirrhosis
    - AIDS_V3 # HIC/AIDS
    - PH            # Portal hypertension
    - RAICU1    # Primary reason for admission to your unit
    - RAICU2    # Secondary reasons for admission to your unit
    - URAICU    # Ultimate primary reason for admission to unit
```

```
numVars:
    HCM: # Height
        microaggregation:
            aggr: 2

datetimeVars:
    DOAH:  # Date of original admission to/attendance at acute hospital'
        microaggregation:
            aggr: 1
        addNoise:
            noise: 2

    DAH: # Date of admission to your hospital
        addNoise:
            noise: 2

    DOAICU:  # Date of original admission to ICU/HDU'
        microaggregation:
            aggr: 1
        addNoise:
            noise: 1
... ...
```

```
conf <- yaml.load_file("../data/test_demo.yaml")
av <- anony.var(conf)
all.var <- c(av$dirv, av$all.vars) # all variables besides non-confidential data
pander(demg.table[, all.var],  style = 'rmarkdown')
```

Table 0.1: Table continues below

pasno	ICNNO	ADNO	NHSNO	TUADNO	DOB	GPCODE	SEX	PCODE	HCM
pas_1	site_1	NA	nhs_1	NA	1988-06-07	GPCODE1	F	NULL	170
pas_2	site_1	NA	nhs_2	NA	1980-12-30	GPCODE1	F	NULL	174
pas_1	site_1	NA	nhs_1	NA	NULL	GPCODE2	F	NULL	170
pas_1	site_1	NA	nhs_1	NA	NULL	GPCODE2	M	NULL	170
pas_1	site_1	NA	nhs_1	NA	NULL	GPCODE4	M	NULL	170

Table 0.2: Table continues below

DOAH	DAH	DOAICU	DUDICU	DOD	DDBSD	DWFRD	TWFRD	DDH
NULL	2014-02-01	NULL	NULL	NULL	NULL	2014-02-01	18:00:00	2014-02-01
NULL	2014-02-01	NULL	NULL	NULL	NULL	2014-02-01	10:00:00	2014-02-01
NULL	2014-02-09	NULL	NULL	NULL	NULL	2014-02-08	18:00:00	2014-02-26
NULL	2014-02-01	NULL	NULL	NULL	NULL	2014-02-01	18:00:00	2014-02-27
NULL	2014-02-01	NULL	NULL	NULL	NULL	2014-02-01	18:00:00	2014-02-01

DUDH	DAICU	DLCCA	DDICU	BPC	AIDS_V3	PH	RAICU1	RAICU2	URAICU
NULL	2014-02-01 10:00	NULL	2014-02-10	NULL	NULL	NULL	1.0	NULL	NULL
NULL	2014-02-01 10:00	NULL	NULL	NULL	NULL	NULL	2.0	NULL	NULL

DUDH	DAICU	DLCCA	DDICU	BPC	AIDS_V3	PH	RAICU1	RAICU2	URAICU
NULL	2014-02-01 10:00	NULL	2014-02-10	NULL	NULL	NULL	1.0	NULL	NULL
NULL	2014-02-01 10:00	NULL	2014-02-10	NULL	NULL	NULL	3.0	NULL	NULL
NULL	2014-02-01 10:00	NULL	2014-02-10	NULL	NULL	NULL	4.0	NULL	NULL