

# Record matching

Adele Tyson

2023-06-23

```
library(janitor)
library(Hmisc)
library(readxl)
library(writexl)
library(reclin2)
library(lubridate)
library(RecordLinkage)
library(dgof) # for statistical testing
library(fdm2id) # for predict that works for kmeans
library(ppclust) # for cmeans
library(tidyverse)

commune_region_lookup <- read_excel("04_Data/Outputs/region_service_commune.xlsx") %>% clean_names()

chile_merged_raw <- read_csv("04_Data/Data_Chile_Merge.csv") %>% clean_names()

chile_merged <- chile_merged_raw %>%
  rename(sex_desc = sex,
         year = agno,
         school_code = rbd,
         school_check_code = dgv_rbd,
         school_name = nom_rbd,
         school_region_code = cod_reg_rbd,
         school_region_name_abr = nom_reg_rbd_a,
         school_province_code = cod_pro_rbd,
         school_commune_code = cod_com_rbd,
         school_commune_name = nom_com_rbd,
         school_dept_code = cod_deprov_rbd,
         school_dept_name = nom_deprov_rbd,
         school_dependency_code = cod_depe, # has categories 1-6, no1 and no2 here are no1 in grouped
         school_dependency_code_grouped = cod_depe2, # has categories 1-5
         school_rurality_code = rural_rbd,
         school_operation_status = estado_estab,
         teaching_code1 = cod_ense, # min = 10, max = 910, eg preschool, special education hearing impa
         teaching_code2 = cod_ense2, # subject matter coding, 1-8
         teaching_code3 = cod_ense3, # age based coding, 1-7
         grade_code1 = cod_grado, # grade of schooling, 1-10, 21-25, 31-34, nests in teaching_code1
         grade_code2 = cod_grado2, # equivalent grade of schooling for adult special education, 1-8, 99
         grade_letter = let_cur, # refers to the class within the grade, close to start of alphabet is
         course_timing = cod_jor, # time of day, morning, afternoon, both, night, no info
         course_type = cod_tip_cur, # 0 = simple course, 1-4 = combined course, 99 = no info
         course_descr = cod_des_cur, # Description of course (TP secondary education only). 0: Does not
         student_id = mrun,
```

```
sex = gen_alu, # 0 = no info, 1 = male, 2 = female
dob = fec_nac_alu_2, # The second one has DD
age_june30 = edad_alu, # age at 30th June 2021
special_needs_status = int_alu, # integrated student indicator, 0 = no, 1 = yes. Mostly no
special_needs_code = cod_int_alu, # ADHD, blindness, etc. 0 = none. 105 = autism, 203 = ADHD.
student_region_code = cod_reg_alu,
student_commune_code = cod_com_alu,
student_commune_name = nom_com_alu,
economic_sector_code = cod_sec,
economic_specialty_code = cod_espe,
economic_branch_code = cod_rama,
economic_profspec_code = cod_men,
teaching_code_new = ens)
```

```
clinical_large_raw <- read_excel("04_Data/dataset_ssas_2015_2021.xlsx") %>% clean_names
#describe(clinical_raw)
```

```
clinical_large <- clinical_large_raw %>%
```

```
  select(c(-procedence, -ethnicity, -education_level, -disability, -foster_care)) %>%
  # Fix the date columns
```

```
  mutate(dob_eng = ifelse(str_detect(date_of_birth, "/"), 1, ifelse(str_detect(date_of_birth, "-"), 0, NA)),
    apt_eng = ifelse(str_detect(date_appointment, "/"), 1, ifelse(str_detect(date_appointment, "-"), 0, NA)),
    dob_day = ifelse(dob_eng == 1, as.integer(str_extract(date_of_birth, "^\\d+")),
      ifelse(dob_eng == 0, as.integer(str_extract(date_of_birth, "^\\d+")), NA)),
    dob_month = ifelse(dob_eng == 1, as.integer(str_extract(date_of_birth, "(?<=/)\\d+(?=/)")),
      ifelse(dob_eng == 0, str_extract(date_of_birth, "(?<=)\\w+(?=)"), NA)),
    dob_year = ifelse(dob_eng == 1, as.integer(str_extract(date_of_birth, "\\d+$")),
      ifelse(dob_eng == 0, as.integer(str_extract(date_of_birth, "\\d+$")) + 2000, NA)),
    dob_month_eng = as.integer(ifelse(dob_month == "ene", 1,
      ifelse(dob_month == "abr", 4,
        ifelse(dob_month == "ago", 8,
          ifelse(dob_month == "sept", 9,
            ifelse(dob_month == "dic", 12, dob_month))
        )
      )
    ),
    dob = make_date(year = dob_year, month = dob_month_eng, day = dob_day),
    apt_day = ifelse(apt_eng == 1, as.integer(str_extract(date_appointment, "^\\d+")),
      ifelse(apt_eng == 0, as.integer(str_extract(date_appointment, "^\\d+")), NA)),
    apt_month = ifelse(apt_eng == 1, as.integer(str_extract(date_appointment, "(?<=/)\\d+(?=/)")),
      ifelse(apt_eng == 0, str_extract(date_appointment, "(?<=)\\w+(?=)"), NA)),
    apt_year = ifelse(apt_eng == 1, as.integer(str_extract(date_appointment, "\\d+$")),
      ifelse(apt_eng == 0, as.integer(str_extract(date_appointment, "\\d+$")) + 2000, NA)),
    apt_month_eng = as.integer(ifelse(apt_month == "ene", 1,
      ifelse(apt_month == "abr", 4,
        ifelse(apt_month == "ago", 8,
          ifelse(apt_month == "sept", 9,
            ifelse(apt_month == "dic", 12, apt_month))
        )
      )
    ),
    apt_date = make_date(year = apt_year, month = apt_month_eng, day = apt_day),
    age_june30 = trunc(time_length(interval(ymd(dob), ymd("2021-06-30")), unit = "year")),
    commune_name = ifelse(comuna == "CHOL CHOL", "CHOLCHOL",
      ifelse(comuna == "CURACAUTIN", "CURACAUTÍN",
        ifelse(comuna == "PITRUFQUEN", "PITRUFQUÉN",
          ifelse(comuna == "PUCON", "PUCÓN",
            ifelse(comuna == "TOLTEN", "TOLTÉN",
              ifelse(comuna == "VILCUN", "VILCÚN",
                NA
              )
            )
          )
        )
      )
    ),
    ses_status = ifelse(socio_economic_level == "FONASA - A", 1,
```



```

clinical_small <- clinical_small %>%
  mutate(intdisab = ifelse(cod_dg_1 %in% intdisab_codes |
                           cod_dg_2 %in% intdisab_codes |
                           cod_dg_3 %in% intdisab_codes, 1, 0)) %>%
  rename("codigo" = "cod_dg_1") %>%
  select(c(-cod_dg_2, -cod_dg_3))

clinical <- rbind(clinical_large, clinical_small)

clinical_communes <- clinical %>% group_by(commune_name) %>% summarise() %>% arrange() %>%
  mutate(commune_in_school_data = ifelse(commune_name %in% unique(chile_merged$student_commune_name), 1, 0))

```

Fixed the date columns because they were in English and Spanish. Redefined the age column to be age at 30th June 2021.

Get one row per person per commune to make matching more efficient. Take the earliest appointment for each person.

```

get.min.na <- function(x) ifelse( !all(is.na(x)), min(x, na.rm = TRUE), NA)
get.max.na <- function(x) ifelse( !all(is.na(x)), max(x, na.rm = TRUE), NA)

patients <- clinical %>%
  group_by(id, gender, dob, commune_name, region_name, ses_status) %>% # Maybe move SES back to here
  summarise(#ses_status = get.min.na(ses_status),
            autism = get.max.na(autism),
            #intdisab = get.max.na(intdisab),
            aut_rank = get.min.na(aut_rank)) %>%
  ungroup() %>%
  rename("student_commune_name" = "commune_name",
        "student_region_name" = "region_name",
        "sex_desc" = "gender") %>%
  rowid_to_column("row_id") %>%
  select(row_id,
        id,
        dob,
        sex_desc,
        student_commune_name,
        autism,
        ses_status,
        #intdisab,
        aut_rank) #, student_region_name) #, count)

```

## `summarise()` has grouped output by 'id', 'gender', 'dob', 'commune\_name',  
## 'region\_name'. You can override using the `.groups` argument.

```
write_xlsx(patients, "04_Data/Outputs/patients.xlsx")
```

```
length(unique(patients$id))
```

```
## [1] 1702
```

```

patients_unique <- patients %>%
  group_by(id) %>%
  summarise(sex_desc = list(sex_desc),
            student_commune_name = list(student_commune_name),
            dob = list(dob),

```

```

    ses_status = list(ses_status))
write_csv(patients_unique, "04_Data/Outputs/patients_unique.csv") # can't write columns containing list

```

NB: there are 1473 unique ID's in patients and it's 1478 rows long, therefore 5 repeated people - probably moved communes.

Are all the records in the small dataset in the big one? No

```

clinical %>% filter(id %in% clinical_small$id)

```

```

## # A tibble: 3,556 x 18
##   id      gender comun~1 healt~2 regio~3 socio~4 ses_s~5 dob      age_j~6
##   <chr>    <chr>  <chr>    <chr>    <chr>    <chr>    <dbl> <date>    <dbl>
## 1 21282495-K Female LONCOCHE Servic~ ARAUC  FONASA~      2 2003-04-16    18
## 2 21282495-K Female LONCOCHE Servic~ ARAUC  FONASA~      2 2003-04-16    18
## 3 21294488-2 Male   VILLARR~ Servic~ ARAUC  Privat~      3 2003-05-15    18
## 4 21294488-2 Male   VILLARR~ Servic~ ARAUC  Privat~      3 2003-05-15    18
## 5 21294488-2 Male   VILLARR~ Servic~ ARAUC  Privat~      3 2003-05-15    18
## 6 21341924-2 Male   PUCÓN    Servic~ ARAUC  FONASA~      2 2003-07-18    17
## 7 21341924-2 Male   PUCÓN    Servic~ ARAUC  FONASA~      2 2003-07-18    17
## 8 21341924-2 Male   PUCÓN    Servic~ ARAUC  FONASA~      2 2003-07-18    17
## 9 21341924-2 Male   PUCÓN    Servic~ ARAUC  FONASA~      2 2003-07-18    17
## 10 21341924-2 Male   PUCÓN    Servic~ ARAUC  FONASA~      2 2003-07-18    17
## # ... with 3,546 more rows, 9 more variables: apt_date <date>, hospital <chr>,
## #   medical_specialty <chr>, type_appointment <chr>, codigo <chr>,
## #   diagnosis <chr>, autism <dbl>, intdisab <dbl>, aut_rank <dbl>, and
## #   abbreviated variable names 1: commune_name, 2: health_service_name,
## #   3: region_name, 4: socio_economic_level, 5: ses_status, 6: age_june30

```

Assume this is because the big clinical dataset only has people with autism, not ADHD.

Only try to link clinical data to records in the schools data for the Southern health service in Araucanía (ARAUC) because that's where the clinical data is from.

```

school <- chile_merged %>%
  rename(commune_name = student_commune_name) %>%
  left_join(commune_region_lookup, by = "commune_name") %>%
  filter(health_service_name == "Servicio de Salud Araucanía Sur") %>% # This should be filtered either
  filter(age_june30 >= 6 & age_june30 <= 18, sex != 0) %>% # Could try without this filter to pick up e
  # filter only the communes represented in the clinical data here?
  mutate(autism = ifelse(special_needs_code == 105, 1, 0),
         #intdisab = 0,
         aut_rank = 1,
         dob = ymd(dob),
         ses_status = ifelse(school_fee == "", NA,
                             ifelse(school_fee == "GRATUITO", 1,
                                     ifelse(school_fee == "$1.000 A $10.000", 2,
                                             ifelse(school_fee == "$10.001 A $25.000", 2,
                                                     ifelse(school_fee == "$25.001 A $50.000", 2,
                                                             ifelse(school_fee == "$50.001 A $100.000", 2,
                                                                     ifelse(school_fee == "MAS DE $100.000",
                                                                                             ifelse(school_fee == "SIN INFORM
filter(autism == 1) %>% # We only want to find additional autism cases in the clinical records, we do
rename(student_region_name = region_name, student_commune_name = commune_name) %>%
select(dob,
       sex_desc,

```

```

        student_commune_name,
        autism,
        ses_status,
        #intdisab,
        aut_rank#,
        #student_region_name
    ) %>%
    rowid_to_column("id")

# Do the commune names align well? Yes
table(sort(unique(patients$student_commune_name, sort(unique(school$student_commune_name)))))

##
##      ALGARROBO      CABO DE HORNO      CARAHUE      CHOLCHOL
##           1              1              47              9
##      CUNCO      CURACAUTÍN      CURARREHUE      DIEGO DE ALMAGRO
##          33              1              12              1
##      FREIRE      GALVARINO      GORBEA      HIJUELAS
##          32              22              21              1
##      LAUTARO      LONCOCHE      LONQUIMAY      MACHALÍ
##         106             89              1              1
##      MELIPEUCO      NUEVA IMPERIAL      PADRE LAS CASAS      PANGUIPULLI
##           5              81             148              1
##      PENCAHUE      PERQUENCO      PICA      PITRUFQUÉN
##           1              19              1              47
##      PUCÓN      QUINTA NORMAL      SAAVEDRA      TEMUCO
##          95              1              14             603
##      TEODORO SCHMIDT      TOCOPILLA      TOLTÉN      VICTORIA
##          12              1              18              1
##      VILCÚN      VILLARRICA
##          60             274

sort(unique(patients$student_commune_name))

## [1] "ALGARROBO"      "CABO DE HORNO"      "CARAHUE"      "CHOLCHOL"
## [5] "CUNCO"          "CURACAUTÍN"        "CURARREHUE"    "DIEGO DE ALMAGRO"
## [9] "FREIRE"         "GALVARINO"         "GORBEA"        "HIJUELAS"
## [13] "LAUTARO"        "LONCOCHE"          "LONQUIMAY"     "MACHALÍ"
## [17] "MELIPEUCO"      "NUEVA IMPERIAL"    "PADRE LAS CASAS" "PANGUIPULLI"
## [21] "PENCAHUE"       "PERQUENCO"         "PICA"          "PITRUFQUÉN"
## [25] "PUCÓN"          "QUINTA NORMAL"     "SAAVEDRA"      "TEMUCO"
## [29] "TEODORO SCHMIDT" "TOCOPILLA"         "TOLTÉN"        "VICTORIA"
## [33] "VILCÚN"         "VILLARRICA"

sort(unique(school$student_commune_name))

## [1] "CARAHUE"      "CHOLCHOL"      "CUNCO"      "CURARREHUE"
## [5] "FREIRE"       "GALVARINO"     "GORBEA"     "LAUTARO"
## [9] "LONCOCHE"     "MELIPEUCO"     "NUEVA IMPERIAL" "PADRE LAS CASAS"
## [13] "PERQUENCO"    "PITRUFQUÉN"    "PUCÓN"      "SAAVEDRA"
## [17] "TEMUCO"       "TEODORO SCHMIDT" "TOLTÉN"     "VILCÚN"
## [21] "VILLARRICA"

```

## Try manual linkage

```
patients_grouped <- patients %>%
  group_by(sex_desc,
            dob,
            student_commune_name) %>%
  summarise(count = n(),
            ids = list(id))
```

## `summarise()` has grouped output by 'sex\_desc', 'dob'. You can override using  
## the `.groups` argument.

```
school_grouped <- school %>%
  group_by(sex_desc,
            dob,
            student_commune_name) %>%
  summarise(count = n(),
            #ids = list(rowid)
            ses = list(ses_status))
```

## `summarise()` has grouped output by 'sex\_desc', 'dob'. You can override using  
## the `.groups` argument.

```
sort(unique(patients$student_commune_name))
```

```
## [1] "ALGARROBO"      "CABO DE HORNOS"  "CARAHUE"         "CHOLCHOL"
## [5] "CUNCO"          "CURACAUTÍN"     "CURARREHUE"      "DIEGO DE ALMAGRO"
## [9] "FREIRE"         "GALVARINO"      "GORBEA"          "HIJUELAS"
## [13] "LAUTARO"        "LONCOCHE"       "LONQUIMAY"       "MACHALÍ"
## [17] "MELIPEUCO"      "NUEVA IMPERIAL" "PADRE LAS CASAS" "PANGUIPULLI"
## [21] "PENCAHUE"       "PERQUENCO"      "PICA"            "PITRUFQUÉN"
## [25] "PUCÓN"          "QUINTA NORMAL"  "SAAVEDRA"        "TEMUCO"
## [29] "TEODORO SCHMIDT" "TOCOPILLA"      "TOLTÉN"          "VICTORIA"
## [33] "VILCÚN"         "VILLARRICA"
```

```
sort(unique(school$student_commune_name))
```

```
## [1] "CARAHUE"        "CHOLCHOL"       "CUNCO"           "CURARREHUE"
## [5] "FREIRE"         "GALVARINO"      "GORBEA"          "LAUTARO"
## [9] "LONCOCHE"       "MELIPEUCO"      "NUEVA IMPERIAL"  "PADRE LAS CASAS"
## [13] "PERQUENCO"      "PITRUFQUÉN"     "PUCÓN"           "SAAVEDRA"
## [17] "TEMUCO"         "TEODORO SCHMIDT" "TOLTÉN"          "VILCÚN"
## [21] "VILLARRICA"
```

```
merged <- merge(school, patients, by = c("sex_desc", "dob", "student_commune_name"), all = TRUE)
merged %>% filter(!is.na(id.x) & !is.na(id.y)) # 85 matches
```

```
##      sex_desc      dob student_commune_name id.x autism.x ses_status.x
## 1   Female 2003-04-16          LONCOCHE    450         1         1
## 2   Female 2003-11-25           TEMUCO    437         1         2
## 3   Female 2005-12-07           TEMUCO    380         1         1
## 4   Female 2006-08-10          LAUTARO    470         1         1
## 5   Female 2006-09-20           FREIRE    109         1         1
## 6   Female 2006-10-10    PADRE LAS CASAS    263         1         1
## 7   Female 2008-05-20           GORBEA    187         1         1
## 8   Female 2008-06-21           TEMUCO    269         1         1
## 9   Female 2009-05-08           TEMUCO     57         1         1
```

## 10	Female	2009-06-22	PUCÓN	332	1	1
## 11	Female	2010-04-27	TEMUCO	426	1	1
## 12	Female	2011-04-20	TEMUCO	173	1	2
## 13	Female	2012-01-31	VILLARRICA	172	1	1
## 14	Female	2012-01-31	VILLARRICA	172	1	1
## 15	Female	2012-04-07	PUCÓN	425	1	1
## 16	Female	2012-05-28	VILCÚN	214	1	1
## 17	Female	2012-06-18	VILLARRICA	41	1	1
## 18	Female	2012-09-13	TEMUCO	104	1	1
## 19	Female	2013-04-20	GALVARINO	296	1	1
## 20	Female	2013-06-19	TEMUCO	267	1	1
## 21	Female	2013-08-30	PADRE LAS CASAS	311	1	1
## 22	Female	2013-12-30	VILLARRICA	190	1	2
## 23	Female	2014-02-15	TEMUCO	105	1	1
## 24	Female	2014-10-09	GORBEA	419	1	1
## 25	Female	2014-10-16	TEMUCO	415	1	2
## 26	Female	2014-11-12	TEMUCO	351	1	1
## 27	Female	2014-12-11	PUCÓN	80	1	1
## 28	Female	2014-12-12	TEMUCO	464	1	1
## 29	Male	2003-01-27	TEMUCO	227	1	1
## 30	Male	2003-03-06	TEMUCO	465	1	1
## 31	Male	2003-06-14	TEMUCO	92	1	1
## 32	Male	2003-06-15	TEMUCO	165	1	1
## 33	Male	2003-06-29	TEMUCO	53	1	1
## 34	Male	2003-08-03	TEMUCO	313	1	1
## 35	Male	2003-10-21	TEMUCO	186	1	1
## 36	Male	2003-12-15	TEMUCO	389	1	1
## 37	Male	2004-03-05	NUEVA IMPERIAL	442	1	1
## 38	Male	2004-03-12	TEMUCO	133	1	1
## 39	Male	2004-07-07	TEMUCO	322	1	2
## 40	Male	2004-09-28	LONCOCHE	216	1	1
## 41	Male	2004-10-01	FREIRE	307	1	1
## 42	Male	2004-11-07	TEMUCO	362	1	1
## 43	Male	2004-12-25	CUNCO	174	1	1
## 44	Male	2005-01-03	TEMUCO	39	1	2
## 45	Male	2005-01-09	TEMUCO	49	1	1
## 46	Male	2005-01-21	TEMUCO	202	1	1
## 47	Male	2005-05-24	TEMUCO	78	1	1
## 48	Male	2005-06-17	TEMUCO	123	1	1
## 49	Male	2005-06-17	TEMUCO	123	1	1
## 50	Male	2005-08-29	TEMUCO	70	1	1
## 51	Male	2005-09-06	TEMUCO	405	1	2
## 52	Male	2006-03-04	TEMUCO	147	1	1
## 53	Male	2006-03-22	TEMUCO	11	1	1
## 54	Male	2006-04-13	PADRE LAS CASAS	301	1	1
## 55	Male	2006-09-09	GALVARINO	434	1	1
## 56	Male	2006-09-19	LAUTARO	219	1	1
## 57	Male	2006-10-06	LAUTARO	448	1	1
## 58	Male	2006-10-10	VILCÚN	478	1	1
## 59	Male	2006-10-27	TEMUCO	247	1	1
## 60	Male	2006-11-02	PADRE LAS CASAS	176	1	2
## 61	Male	2006-11-06	TEMUCO	471	1	2
## 62	Male	2006-11-06	TEMUCO	471	1	2
## 63	Male	2007-01-08	CARAHUE	319	1	1



## 64	Male	2007-01-23	VILLARRICA	363	1	1
## 65	Male	2007-02-13	TEMUCO	235	1	1
## 66	Male	2007-03-22	LAUTARO	265	1	1
## 67	Male	2007-04-09	PADRE LAS CASAS	31	1	1
## 68	Male	2007-04-25	LAUTARO	336	1	1
## 69	Male	2007-05-11	TEMUCO	355	1	1
## 70	Male	2007-06-16	PITRUFQUÉN	358	1	1
## 71	Male	2007-08-20	PITRUFQUÉN	237	1	1
## 72	Male	2007-11-06	VILLARRICA	295	1	1
## 73	Male	2007-12-28	LONCOCHE	130	1	1
## 74	Male	2008-01-28	NUEVA IMPERIAL	44	1	1
## 75	Male	2008-03-05	PUCÓN	420	1	1
## 76	Male	2008-03-14	TEMUCO	408	1	1
## 77	Male	2008-03-25	TEMUCO	289	1	1
## 78	Male	2008-03-25	TEMUCO	289	1	1
## 79	Male	2008-05-20	PADRE LAS CASAS	100	1	1
## 80	Male	2008-06-18	VILCÚN	55	1	1
## 81	Male	2008-08-24	SAAVEDRA	158	1	1
## 82	Male	2008-10-10	TEMUCO	112	1	2
## 83	Male	2008-10-22	VILLARRICA	72	1	1
## 84	Male	2008-10-22	VILLARRICA	72	1	1
## 85	Male	2008-11-22	NUEVA IMPERIAL	467	1	1
## 86	Male	2008-12-06	LAUTARO	22	1	1
## 87	Male	2008-12-21	TEMUCO	394	1	1
## 88	Male	2008-12-29	TEMUCO	93	1	1
## 89	Male	2009-01-07	LAUTARO	361	1	1
## 90	Male	2009-01-12	TEMUCO	26	1	1
## 91	Male	2009-02-13	PUCÓN	3	1	1
## 92	Male	2009-02-26	LONCOCHE	168	1	1
## 93	Male	2009-04-23	LONCOCHE	314	1	1
## 94	Male	2009-04-23	LONCOCHE	314	1	1
## 95	Male	2009-08-05	VILLARRICA	60	1	1
## 96	Male	2009-08-05	VILLARRICA	60	1	1
## 97	Male	2009-08-14	PUCÓN	252	1	1
## 98	Male	2009-08-14	PUCÓN	252	1	1
## 99	Male	2009-08-29	TEMUCO	159	1	2
## 100	Male	2009-10-01	TEMUCO	328	1	1
## 101	Male	2009-10-26	TEMUCO	341	1	1
## 102	Male	2010-01-02	FREIRE	272	1	1
## 103	Male	2010-01-25	PADRE LAS CASAS	73	1	2
## 104	Male	2010-02-21	LONCOCHE	180	1	1
## 105	Male	2010-02-26	TEODORO SCHMIDT	213	1	1
## 106	Male	2010-03-07	LAUTARO	242	1	1
## 107	Male	2010-03-16	GORBEA	246	1	1
## 108	Male	2010-05-20	VILLARRICA	396	1	1
## 109	Male	2010-06-07	TEMUCO	476	1	1
## 110	Male	2010-06-08	NUEVA IMPERIAL	292	1	1
## 111	Male	2010-07-21	CHOLCHOL	194	1	1
## 112	Male	2010-07-28	FREIRE	382	1	1
## 113	Male	2010-08-29	VILLARRICA	365	1	1
## 114	Male	2010-09-13	PADRE LAS CASAS	312	1	1
## 115	Male	2010-10-12	TEMUCO	201	1	1
## 116	Male	2010-12-09	PUCÓN	346	1	1
## 117	Male	2010-12-09	TEMUCO	107	1	1

## 118	Male	2011-01-13	VILLARRICA	18	1	2
## 119	Male	2011-01-24	TEMUCO	87	1	1
## 120	Male	2011-02-11	CUNCO	368	1	1
## 121	Male	2011-02-22	TEMUCO	139	1	1
## 122	Male	2011-03-03	LAUTARO	228	1	1
## 123	Male	2011-04-13	VILLARRICA	275	1	1
## 124	Male	2011-04-13	VILLARRICA	275	1	1
## 125	Male	2011-06-13	TEMUCO	203	1	1
## 126	Male	2011-07-02	LAUTARO	475	1	1
## 127	Male	2011-08-02	CARAHUE	113	1	1
## 128	Male	2011-09-06	TEODORO SCHMIDT	229	1	1
## 129	Male	2011-09-08	TEMUCO	277	1	1
## 130	Male	2011-10-27	TEODORO SCHMIDT	283	1	1
## 131	Male	2011-11-10	FREIRE	300	1	1
## 132	Male	2011-11-12	PADRE LAS CASAS	278	1	1
## 133	Male	2012-01-11	PUCÓN	290	1	1
## 134	Male	2012-01-11	PUCÓN	290	1	1
## 135	Male	2012-03-06	VILLARRICA	243	1	1
## 136	Male	2012-03-12	TEMUCO	261	1	1
## 137	Male	2012-04-16	TEMUCO	472	1	1
## 138	Male	2012-05-29	TEMUCO	8	1	1
## 139	Male	2012-06-01	PADRE LAS CASAS	66	1	1
## 140	Male	2012-06-02	TEMUCO	141	1	1
## 141	Male	2012-06-25	GALVARINO	183	1	1
## 142	Male	2012-07-08	TEMUCO	315	1	1
## 143	Male	2012-07-16	GALVARINO	152	1	1
## 144	Male	2012-07-29	VILCÚN	16	1	1
## 145	Male	2012-09-07	TEMUCO	293	1	1
## 146	Male	2012-09-21	CUNCO	264	1	1
## 147	Male	2012-10-13	VILLARRICA	45	1	1
## 148	Male	2012-10-13	VILLARRICA	45	1	1
## 149	Male	2012-10-18	VILLARRICA	392	1	1
## 150	Male	2012-11-03	LAUTARO	443	1	1
## 151	Male	2012-11-05	TEMUCO	447	1	1
## 152	Male	2012-12-10	PITRUFQUÉN	304	1	NA
## 153	Male	2012-12-25	PADRE LAS CASAS	29	1	1
## 154	Male	2013-01-26	GORBEA	385	1	1
## 155	Male	2013-01-30	PITRUFQUÉN	97	1	1
## 156	Male	2013-02-12	TEMUCO	366	1	1
## 157	Male	2013-02-25	GORBEA	294	1	1
## 158	Male	2013-02-27	NUEVA IMPERIAL	24	1	1
## 159	Male	2013-03-24	VILLARRICA	386	1	1
## 160	Male	2013-04-23	TOLTÉN	350	1	1
## 161	Male	2013-05-20	TEMUCO	280	1	1
## 162	Male	2013-05-23	LAUTARO	189	1	1
## 163	Male	2013-05-30	VILLARRICA	469	1	1
## 164	Male	2013-07-07	VILCÚN	338	1	1
## 165	Male	2013-10-16	VILCÚN	111	1	1
## 166	Male	2013-10-23	PITRUFQUÉN	324	1	1
## 167	Male	2013-11-05	VILLARRICA	211	1	1
## 168	Male	2013-11-05	VILLARRICA	211	1	1
## 169	Male	2013-11-14	TEMUCO	71	1	1
## 170	Male	2014-02-19	TEMUCO	326	1	1
## 171	Male	2014-02-19	TEMUCO	326	1	1

## 172	Male	2014-02-19	TEMUCO	451	1	1
## 173	Male	2014-02-19	TEMUCO	451	1	1
## 174	Male	2014-04-17	TEMUCO	335	1	1
## 175	Male	2014-04-21	TEMUCO	129	1	1
## 176	Male	2014-05-06	VILLARRICA	271	1	1
## 177	Male	2014-05-17	CUNCO	125	1	1
## 178	Male	2014-05-20	TEMUCO	287	1	1
## 179	Male	2014-05-24	LONCOCHE	407	1	1
## 180	Male	2014-06-02	TEMUCO	162	1	1
## 181	Male	2014-06-16	TEMUCO	77	1	1
## 182	Male	2014-07-07	TEMUCO	116	1	1
## 183	Male	2014-08-30	GALVARINO	6	1	1
## 184	Male	2014-09-06	TEMUCO	145	1	1
## 185	Male	2014-09-06	TEMUCO	145	1	1
## 186	Male	2014-09-12	LONCOCHE	310	1	1
## 187	Male	2014-09-12	LONCOCHE	310	1	1
## 188	Male	2014-10-07	TEMUCO	98	1	1
## 189	Male	2014-10-07	TEMUCO	98	1	1
## 190	Male	2014-10-07	TEMUCO	118	1	1
## 191	Male	2014-10-07	TEMUCO	118	1	1
## 192	Male	2014-10-28	TEMUCO	15	1	1
## 193	Male	2014-11-02	TEMUCO	399	1	1
## 194	Male	2014-11-16	LAUTARO	91	1	1
## 195	Male	2014-11-19	PUCÓN	357	1	1
## 196	Male	2014-12-29	PADRE LAS CASAS	456	1	2
## 197	Male	2015-01-03	VILLARRICA	353	1	1
## 198	Male	2015-01-19	VILCÚN	157	1	1
## 199	Male	2015-01-25	PADRE LAS CASAS	270	1	1
## 200	Male	2015-02-02	TEODORO SCHMIDT	466	1	1
## 201	Male	2015-03-06	NUEVA IMPERIAL	458	1	1
## 202	Male	2015-03-10	GALVARINO	181	1	1
## 203	Male	2015-03-11	TEMUCO	387	1	1
## 204	Male	2015-03-13	TEMUCO	256	1	1
## 205	Male	2015-05-02	TEMUCO	376	1	1
##	aut_rank.x	row_id	id.y	autism.y	ses_status.y	aut_rank.y
## 1	1	21	21282495-K	1	2	1
## 2	1	81	21449127-3	1	2	1
## 3	1	296	21994583-3	1	1	1
## 4	1	363	22183641-3	1	2	1
## 5	1	374	22213761-6	1	2	1
## 6	1	380	22234827-7	1	2	1
## 7	1	571	22724176-4	1	1	1
## 8	1	583	22752332-8	1	1	1
## 9	1	697	23021556-1	1	2	1
## 10	1	707	23054104-3	1	2	1
## 11	1	824	23310188-5	1	2	1
## 12	1	974	23624343-5	1	2	1
## 13	1	1075	23860402-8	1	2	1
## 14	1	1074	23860402-8	1	1	1
## 15	1	1092	23917587-2	1	2	1
## 16	1	1121	23959967-2	1	3	1
## 17	1	1133	23987283-2	1	1	1
## 18	1	1177	24064290-5	1	2	1
## 19	1	1268	24249709-0	1	1	1

## 20	1	1291	24307066-K	1	2	1
## 21	1	1319	24396036-3	1	2	1
## 22	1	1354	24495784-6	1	2	1
## 23	1	1371	24539730-5	1	1	1
## 24	1	1465	24763669-2	1	2	1
## 25	1	1469	24771215-1	1	2	1
## 26	1	1476	24797188-2	1	1	1
## 27	1	1488	24825751-2	1	2	1
## 28	1	1485	24824555-7	1	2	1
## 29	1	46	21338851-7	1	2	1
## 30	1	14	21251752-6	1	1	1
## 31	1	37	21319146-2	1	1	1
## 32	1	38	21319994-3	1	2	1
## 33	1	43	21332821-2	1	1	1
## 34	1	49	21354095-5	1	2	1
## 35	1	76	21417599-1	1	1	1
## 36	1	87	21464033-3	1	2	1
## 37	1	102	21520695-5	1	2	1
## 38	1	108	21543736-1	1	2	1
## 39	1	138	21619878-6	1	2	1
## 40	1	162	21670184-4	1	2	1
## 41	1	165	21679874-0	1	2	1
## 42	1	174	21700914-6	1	1	1
## 43	1	192	21737462-6	1	2	1
## 44	1	196	21748664-5	1	2	1
## 45	1	198	21750199-7	1	2	1
## 46	1	202	21759050-7	1	1	1
## 47	1	243	21859877-3	1	2	1
## 48	1	245	21862073-6	1	2	1
## 49	1	246	21867880-7	1	2	1
## 50	1	263	21921022-1	1	1	1
## 51	1	265	21925304-4	1	2	1
## 52	1	318	22065375-7	1	1	1
## 53	1	323	22079654-K	1	2	1
## 54	1	331	22095157-K	1	2	1
## 55	1	368	22204715-3	1	1	1
## 56	1	373	22211545-0	1	2	1
## 57	1	379	22226291-7	1	2	1
## 58	1	381	22237373-5	1	1	1
## 59	1	387	22245810-2	1	1	1
## 60	1	388	22249166-5	1	2	1
## 61	1	392	22253904-8	1	2	1
## 62	1	391	22253752-5	1	2	1
## 63	1	415	22300065-7	1	2	1
## 64	1	419	22312842-4	1	2	1
## 65	1	430	22327040-9	1	2	1
## 66	1	439	22356979-K	1	2	1
## 67	1	446	22370213-9	1	2	1
## 68	1	453	22386477-5	1	2	1
## 69	1	456	22395859-1	1	2	1
## 70	1	461	22426890-4	1	2	1
## 71	1	486	22491627-2	1	2	1
## 72	1	502	22549846-6	1	2	1
## 73	1	521	22592217-9	1	1	1

## 74	1	532	22637968-1	1	2	1
## 75	1	539	22663017-1	1	2	1
## 76	1	547	22670294-6	1	1	1
## 77	1	551	22678488-8	1	2	1
## 78	1	550	22678488-8	1	1	1
## 79	1	570	22723986-7	1	1	1
## 80	1	585	22755037-6	1	2	1
## 81	1	605	22805100-4	1	1	1
## 82	1	624	22838644-8	1	2	1
## 83	1	634	22852889-7	1	1	1
## 84	1	635	22852889-7	1	2	1
## 85	1	641	22881315-K	1	1	1
## 86	1	645	22891576-9	1	2	1
## 87	1	650	22901266-5	1	2	1
## 88	1	655	22907807-0	1	2	1
## 89	1	657	22915922-4	1	2	1
## 90	1	658	22920380-0	1	2	1
## 91	1	669	22945155-3	1	2	1
## 92	1	675	22958693-9	1	1	1
## 93	1	692	23006189-0	1	1	1
## 94	1	693	23006189-0	1	2	1
## 95	1	720	23093195-K	1	3	1
## 96	1	719	23093195-K	1	2	1
## 97	1	722	23099554-0	1	3	1
## 98	1	721	23099554-0	1	2	1
## 99	1	726	23111138-7	1	2	1
## 100	1	737	23136875-2	1	2	1
## 101	1	753	23157810-2	1	2	1
## 102	1	785	23216852-8	1	1	1
## 103	1	794	23233498-3	1	2	1
## 104	1	802	23258114-K	1	1	1
## 105	1	809	23266559-9	1	1	1
## 106	1	807	23263729-3	1	1	1
## 107	1	812	23273376-4	1	2	1
## 108	1	833	23330047-0	1	2	1
## 109	1	839	23343300-4	1	2	1
## 110	1	840	23346792-8	1	1	1
## 111	1	855	23378083-9	1	2	1
## 112	1	858	23386130-8	1	1	1
## 113	1	871	23410879-4	1	2	1
## 114	1	874	23423713-6	1	2	1
## 115	1	883	23448369-2	1	2	1
## 116	1	913	23506849-4	1	1	1
## 117	1	908	23501831-4	1	2	1
## 118	1	927	23534842-K	1	2	1
## 119	1	932	23543378-8	1	1	1
## 120	1	941	23559600-8	1	2	1
## 121	1	945	23567468-8	1	2	1
## 122	1	949	23574393-0	1	1	1
## 123	1	975	23625011-3	1	1	1
## 124	1	976	23625011-3	1	2	1
## 125	1	992	23667140-2	1	2	1
## 126	1	996	23683414-K	1	2	1
## 127	1	1005	23713649-7	1	1	1

## 128	1	1014	23737580-7	1	2	1
## 129	1	1016	23740506-4	1	2	1
## 130	1	1039	23785220-6	1	1	1
## 131	1	1044	23794254-K	1	1	1
## 132	1	1045	23795374-6	1	1	1
## 133	1	1064	23843993-0	1	1	1
## 134	1	1065	23843993-0	1	2	1
## 135	1	1086	23896217-K	1	2	1
## 136	1	1087	23900150-5	1	3	1
## 137	1	1099	23929914-8	1	1	1
## 138	1	1124	23967787-8	1	1	1
## 139	1	1125	23968562-5	1	2	1
## 140	1	1126	23969130-7	1	2	1
## 141	1	1136	23994954-1	1	1	1
## 142	1	1143	24005478-7	1	1	1
## 143	1	1148	24014350-K	1	2	1
## 144	1	1155	24026293-2	1	2	1
## 145	1	1171	24058690-8	1	1	1
## 146	1	1184	24073081-2	1	2	1
## 147	1	1193	24092534-6	1	2	1
## 148	1	1194	24092534-6	1	3	1
## 149	1	1195	24093718-2	1	2	1
## 150	1	1211	24121753-1	1	1	1
## 151	1	1199	24107434-K	1	2	1
## 152	1	1217	24139241-4	1	2	1
## 153	1	1224	24152537-6	1	1	1
## 154	1	1235	24182326-1	1	2	1
## 155	1	1234	24180190-K	1	2	1
## 156	1	1238	24190413-K	1	2	1
## 157	1	1245	24204418-5	1	2	1
## 158	1	1250	24210618-0	1	1	1
## 159	1	1255	24230863-8	1	1	1
## 160	1	1269	24251559-5	1	1	1
## 161	1	1279	24281126-7	1	1	1
## 162	1	1281	24286764-5	1	1	1
## 163	1	1284	24291235-7	1	2	1
## 164	1	1299	24324822-1	1	1	1
## 165	1	1328	24417134-6	1	2	1
## 166	1	1332	24426016-0	1	2	1
## 167	1	1336	24447255-9	1	2	1
## 168	1	1335	24447255-9	1	1	1
## 169	1	1559	24989671-3	1	2	1
## 170	1	1372	24540592-8	1	2	1
## 171	1	1373	24540729-7	1	2	1
## 172	1	1372	24540592-8	1	2	1
## 173	1	1373	24540729-7	1	2	1
## 174	1	1394	24598516-9	1	1	1
## 175	1	1396	24599994-1	1	3	1
## 176	1	1401	24612954-1	1	2	1
## 177	1	1408	24627145-3	1	2	1
## 178	1	1410	24628839-9	1	1	1
## 179	1	1411	24629598-0	1	1	1
## 180	1	1415	24636672-1	1	2	1
## 181	1	1422	24653340-7	1	2	1

## 182	1	1442	24703686-5	1	1	1
## 183	1	1451	24729625-5	1	1	1
## 184	1	1455	24737432-9	1	1	1
## 185	1	1460	24743808-4	1	2	1
## 186	1	1459	24743802-5	1	1	1
## 187	1	1458	24743750-9	1	1	1
## 188	1	1464	24761476-1	1	2	1
## 189	1	1467	24766324-K	1	2	1
## 190	1	1464	24761476-1	1	2	1
## 191	1	1467	24766324-K	1	2	1
## 192	1	1471	24786561-6	1	1	1
## 193	1	1470	24786417-2	1	2	1
## 194	1	1478	24801153-K	1	2	1
## 195	1	1479	24806938-4	1	2	1
## 196	1	1500	24842142-8	1	3	1
## 197	1	1503	24851058-7	1	1	1
## 198	1	1509	24867787-2	1	3	1
## 199	1	1514	24878818-6	1	1	1
## 200	1	1518	24887657-3	1	1	1
## 201	1	1532	24923775-2	1	1	1
## 202	1	1531	24922934-2	1	2	1
## 203	1	1533	24926007-K	1	2	1
## 204	1	1535	24927693-6	1	1	1
## 205	1	1552	24972952-3	1	2	1

All the students that can't be uniquely identified (6) are males in Temuco and SES status doesn't help distinguish them.

## Probabilistic record linkage

<https://rpubs.com/ahmademad/RecordLinkage> <https://www.bristol.ac.uk/media-library/sites/cmm/migrated/documents/problinkage.pdf> <https://cran.r-project.org/web/packages/diyar/vignettes/links.html>

Mismatch on ses is slightly higher weighted than match on everything. Unclear why and doesn't occur for `epiWeights()` below.

```
# Try supplying error information. Works better when sex_desc and dob are both in blocking as otherwise
# Still quick for whole school dataset
a2 <- compare.linkage(school,
  #select(school, -ses_status),
  select(patients, -row_id),
  #select(patients, -ses_status),
  blockfld = c("sex_desc", "dob"), # Block on sex and dob because we really want them
  #blockfld = FALSE,
  phonetic = FALSE,
  strcmp = c(2), # Do string comparison on DOB
  exclude = c(1) # Exclude the id column in both datasets
)
a2_pairs <- a2$pairs # Issue with ses matching here
b2 <- epiWeights(a2, e = c(0.01, # Default for DOB
  0.01, # Default for sex
  0.01, # Default for commune because we want a good match
  0.01, # Keep small so autism in clinical (not intellectual disability) is precise
  0.6, # Have more error for ses_status because it is loosely defined
  #0.3, # Allow more mismatch intellectual disability status so that autism matches
```

```

))
summary(b2)

```

```

##
## Linkage Data Set
##
## 487 records in data set 1
## 1767 records in data set 2
## 315 record pairs
##
## 0 matches
## 0 non-matches
## 315 pairs with unknown status
##
##
## Weight distribution:
##
## [0.55,0.6] [0.6,0.65] [0.65,0.7] [0.7,0.75] [0.75,0.8] [0.8,0.85] [0.85,0.9]
##          10          100           0           0           0           0          205

```

```

allPairs2 <- getPairs(b2)
head(allPairs2, n = 20)

```

```

##      id      id      dob sex_desc student_commune_name autism ses_status
## 1    450      450 2003-04-16  Female          LONCOCHE      1          1
## 2    21 21282495-K 2003-04-16  Female          LONCOCHE      1          2
## 3
## 4    437      437 2003-11-25  Female          TEMUCO       1          2
## 5    81 21449127-3 2003-11-25  Female          TEMUCO       1          2
## 6
## 7    380      380 2005-12-07  Female          TEMUCO       1          1
## 8    296 21994583-3 2005-12-07  Female          TEMUCO       1          1
## 9
## 10   470      470 2006-08-10  Female          LAUTARO      1          1
## 11   363 22183641-3 2006-08-10  Female          LAUTARO      1          2
## 12
## 13   109      109 2006-09-20  Female          FREIRE       1          1
## 14   374 22213761-6 2006-09-20  Female          FREIRE       1          2
## 15
## 16   263      263 2006-10-10  Female    PADRE LAS CASAS      1          1
## 17   380 22234827-7 2006-10-10  Female    PADRE LAS CASAS      1          2
## 18
## 19   187      187 2008-05-20  Female          GORBEA       1          1
## 20   571 22724176-4 2008-05-20  Female          GORBEA       1          1
##      aut_rank      Weight
## 1            1
## 2            1 0.8505747
## 3
## 4            1
## 5            1 0.8505747
## 6
## 7            1
## 8            1 0.8505747

```



```
## 9
## 10      1
## 11      1 0.8505747
## 12
## 13      1
## 14      1 0.8505747
## 15
## 16      1
## 17      1 0.8505747
## 18
## 19      1
## 20      1 0.8505747
```

```
classifyPairs2 <- emClassify(b2, threshold.upper = 1, threshold.lower = 0.8)
a2_pairs$weight <- classifyPairs2$Wdata
a2_pairs$pred <- classifyPairs2$prediction
```

```
a2_pairs_clean <- a2_pairs %>%
  rename(".x" = id1, ".y" = id2) %>%
  select(-is_match)
```

```
finalPairs2 <- getPairs(b2, max.weight = 1, min.weight = 0, single.rows = TRUE) # Take them all when bl
```

```
#kmeansRes2 <- classifyUnsup(a2, method = "kmeans")
#a2_pairs$pred <- kmeansRes2$prediction
# Works but prioritises ses over commune and doesn't use epiWeights found above so not that useful.
```

finalPairs2 is the same size as finalPairs and probably contains the same matches but was much quicker to run because of the blocking. Assume in kmeansRes2, N = not a match, L = likely a match.

```
# reclin has a 1-1 matching fuction so regenerate the pairs using reclin so they're a pairs
# type object and can be passed to select_n_to_m
```

```
pairs <- pair_blocking(school, patients, on = c("sex_desc", "dob")) %>%
  mutate(student_commune_name = (school$student_commune_name[.x] == patients$student_commune_name[.y])
  #ses = get_num_diff(school$ses_status[.x], patients$ses_status[.y])$val
  ) %>%
  left_join(a2_pairs_clean, by = c(".x", ".y")) %>%
  select(c(-student_commune_name.x)) %>%
  rename("student_commune_name" = "student_commune_name.y")
```

```
matches <- select_n_to_m(pairs, threshold = 0.5, score = "weight", n = 1, m = 1, var = "match") %>%
  filter(match == TRUE) %>%
  rename("id" = ".x")
```

```
# In the case of tied fully-perfect matches, might be better to choose the clinical record with autism
```

```
# Now add the matched clinical records to the school records
```

```
school_matched <- school %>%
  left_join(matches, by = "id") %>%
  rename(id.school = id,
         dob.school = dob.x,
         sex_desc.school = sex_desc.x,
         student_commune_name.school = student_commune_name.x,
```

```

    ses_status.school = ses_status.x,
    row_id = .y,
    dob.matched = dob.y,
    sex_desc.matched = sex_desc.y,
    student_commune_name.matched = student_commune_name.y,
    ses_status.matched = ses_status.y) %>%
select(c(-pred, -match)) %>%
left_join(patients, by = "row_id") %>%
rename(id.patient = row_id,
       patient_id = id,
       dob.patient = dob,
       sex_desc.patient = sex_desc,
       student_commune_name.patient = student_commune_name,
       ses_status.patient = ses_status) %>%
select(id.school, id.patient, patient_id,
       dob.school, dob.patient, dob.matched,
       sex_desc.school, sex_desc.patient, sex_desc.matched,
       student_commune_name.school, student_commune_name.patient, student_commune_name.matched,
       ses_status.school, ses_status.patient, ses_status.matched,
       weight) %>%
arrange(desc(weight))

write_csv(school_matched, "04_Data/Outputs/school_matched.csv")

school_matched_yes <- school_matched %>% filter(!is.na(weight))
school_matched_no <- school_matched %>% filter(is.na(weight))

school_matched_small <- school_matched %>%
  mutate(matched = ifelse(is.na(patient_id), 0, 1),
         sex.school = ifelse(sex_desc.school == "Male", 1, ifelse(sex_desc.school == "Female", 2, NA)))
select(id.school, dob.school, sex_desc.school, sex.school, student_commune_name.school, ses_status.school)

```

Consider whether the matched and unmatched school records are different

We hope they are not different

```

library(coin)

## Warning: package 'coin' was built under R version 4.2.3
##
## Attaching package: 'coin'
## The following objects are masked from 'package:arules':
##
##   size, support
# Explore distributions across matched/non-matched
# Sex
data1 <- school_matched_small %>% filter(matched == 1) %>% select(sex.school)
data2 <- school_matched_small %>% filter(matched == 0) %>% select(sex.school)
#hist(data1$sex.school)
#hist(data2$sex.school)
data1 %>% group_by(sex.school) %>% summarise(count = n()) %>% mutate(freq = count/sum(count))

## # A tibble: 2 x 3

```

```

##   sex.school count   freq
##       <dbl> <int> <dbl>
## 1         1    207 0.870
## 2         2     31 0.130

data2 %>% group_by(sex.school) %>% summarise(count = n()) %>% mutate(freq = count/sum(count))

## # A tibble: 2 x 3
##   sex.school count   freq
##       <dbl> <int> <dbl>
## 1         1    219 0.880
## 2         2     30 0.120

ks.sex <- ks.test(data1$sex.school, data2$sex.school, alternative = "two.sided", simulate.p.value = TRUE)

## Warning in ks.test(data1$sex.school, data2$sex.school, alternative =
## "two.sided", : cannot compute correct p-values with ties

ks.sex

##
## Two-sample Kolmogorov-Smirnov test
##
## data:  data1$sex.school and data2$sex.school
## D = 0.0097702, p-value = 1
## alternative hypothesis: two-sided

pt.sex <- oneway_test(sex.school ~ as.factor(matched), data = school_matched_small, distribution = appropr
#confint(pt.sex)

ks.ses <- ks.test(data1$ses_status.school, data2$ses_status.school, alternative = "two.sided", simulate.p.value = TRUE)

## Warning in ks.test(data1$ses_status.school, data2$ses_status.school,
## alternative = "two.sided", : cannot compute correct p-values with ties

ks.ses

##
## Two-sample Kolmogorov-Smirnov test
##
## data:  data1$ses_status.school and data2$ses_status.school
## D = 0.08118, p-value = 0.4067
## alternative hypothesis: two-sided

# SES
#data1 <- school_matched_yes %>% select(ses_status.school)
#data2 <- school_matched_no %>% select(ses_status.school)
#hist(data1$ses_status.school, breaks = 10)
#hist(data2$ses_status.school, breaks = 10)
data1 %>% group_by(ses_status.school) %>% summarise(count = n()) %>% mutate(freq = count/sum(count))

## # A tibble: 3 x 3
##   ses_status.school count   freq
##       <dbl> <int> <dbl>
## 1         1    213 0.895
## 2         2     22 0.0924
## 3        NA      3 0.0126

```

```

data2 %>% group_by(ses_status.school) %>% summarise(count = n()) %>% mutate(freq = count/sum(count))

## # A tibble: 3 x 3
##   ses_status.school count   freq
##           <dbl> <int> <dbl>
## 1             1    203 0.815
## 2             2     43 0.173
## 3            NA      3 0.0120

# Kolmogorov tests with permutation distributions

set.seed(123)
nPerm <- 2000
ks_pvals <- data.frame(sex = numeric(nPerm),
                      ses = numeric(nPerm))

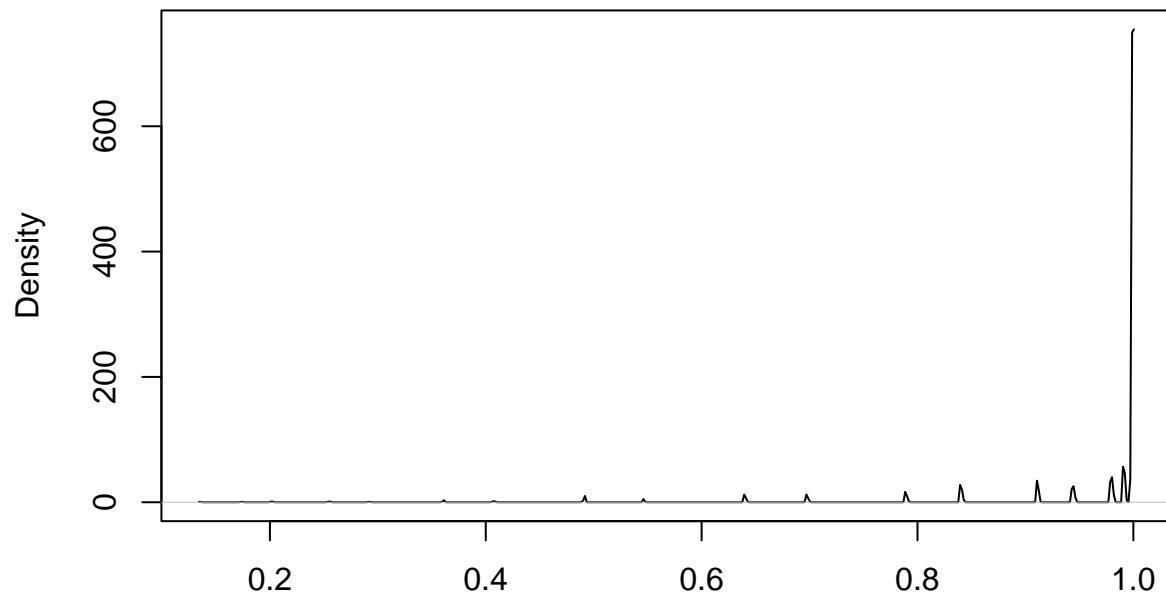
school_matched_small_perm <- school_matched_small

for (i in 1:nPerm) {
  #print(i)
  school_matched_small_perm$matched <- school_matched_small$matched[sample(nrow(school_matched_small))]
  datayes <- school_matched_small_perm %>% filter(matched == 1)
  datano <- school_matched_small_perm %>% filter(matched == 0)
  ks.sex <- ks.test(datayes$sex.school, datano$sex.school, alternative = "two.sided")
  ks.ses <- ks.test(datayes$ses_status.school, datano$ses_status.school, alternative = "two.sided")
  ks_pvals$sex[i] <- ks.sex$p.value
  ks_pvals$ses[i] <- ks.ses$p.value
}

plot(density(ks_pvals$sex))

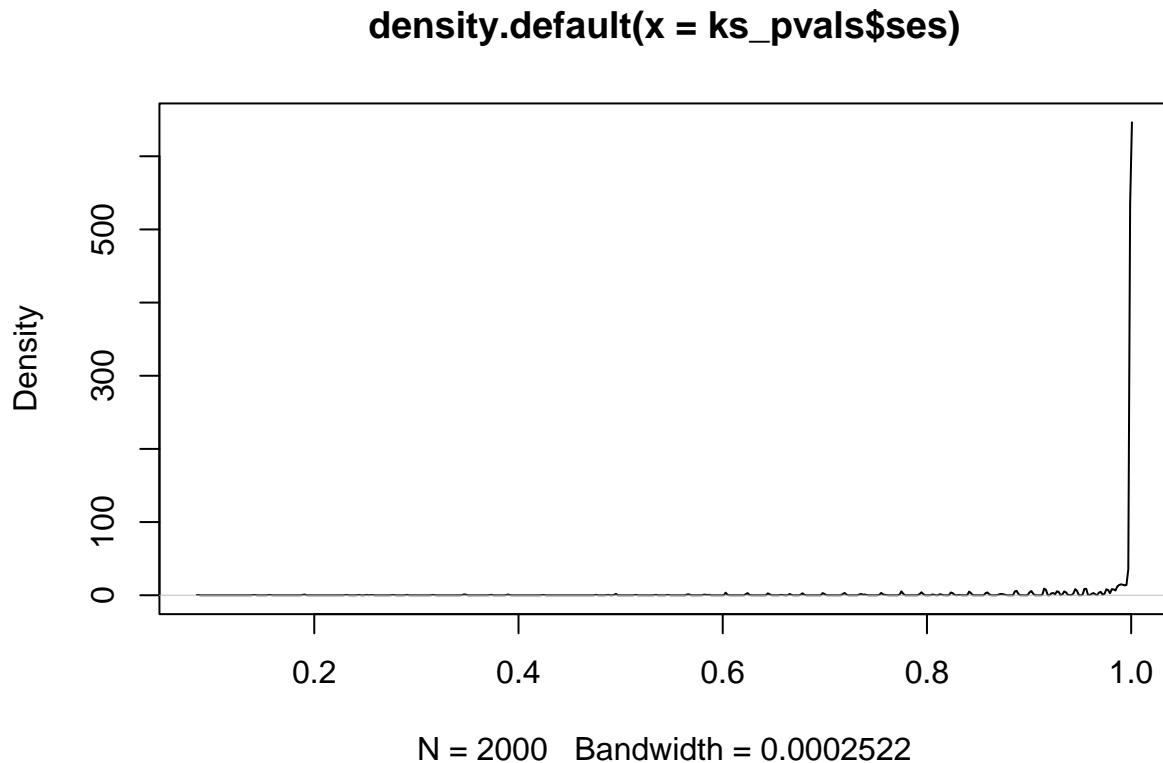
```

**density.default(x = ks\_pvals\$sex)**



N = 2000 Bandwidth = 0.0001866

```
#hist(ks_pvals$sex, breaks = 30)  
plot(density(ks_pvals$ses))
```



```
#hist(ks_pvals$ses, breaks = 30)
```

Our matched/non-matched are not different by sex (p-value in Kolmog is same as most of distribution of permuted pvals) but are different by ses status

Probably not suitable to use K-S because data are not continuous. Try Chi2 instead. Or simulated K-S to choose our own p-value threshold.

Now analyse the matched and unmatched school records and see if they are fundamentally different. ?Cohen test?

Then quantify clinical records for ARAUC Sur that haven't been matched.

**Dumping ground, don't use below here.**

## Record linkage using machine learning

Try linkage using ML, as done by Jan van der Laan here [https://cran.r-project.org/web/packages/reclin2/vignettes/record\\_linkage\\_using\\_machine\\_learning.html](https://cran.r-project.org/web/packages/reclin2/vignettes/record_linkage_using_machine_learning.html)

In reclin2 package, use ?identical() to see available matching algorithms.

The Jaro-Winkler distance is a string metric for measuring the edit distance between two sequences. It is a variant of the Jaro distance metric proposed by William E. Winkler in 1990 <sup>1</sup>. The Jaro-Winkler distance uses a prefix scale which gives more favorable ratings to strings that match from the beginning for a set prefix length. The higher the Jaro-Winkler distance for two strings is, the less similar the strings are. The score is normalized such that 0 means an exact match and 1 means there is no similarity <sup>1</sup>.

Need to explore different comparator algorithms. Currently it's exact match. Would be good to do communes that are neighbours and ages off by 1.

## **Try bayesian linkage?**

Follow Thomas Stringham <https://arxiv.org/pdf/2003.04238.pdf> who followed Sadinle <https://arxiv.org/abs/1601.06630> Not doing this as limited value when not matching strings.