100 Years of Social Work Research: A Data Science Perspective

Overview of data

The original data were from a search of PsychInfo using Ebsco Host platform (December 23, 2014). The following search operators and limiters were used:

- SO "social work" OR SO "social welfare" OR SO "social casework" OR SO "social services"
- Limiters - Document Type: Journal Article
- Search modes - Boolean/Phrase Interface - EBSCOhost Research Databases
- Search Screen - Advanced Search
- Database - PsycINFO

The search results were exported in a *generic bibliographic format*, which is an unstructured text (*.txt) file.The text file was processed using the `BibWrangleR` function created by the first author.

Initialize OS-X workspace and functions for data wrangling

This section processes raw data. This section of code is executed only one time to transform raw text data into an analyzable format. When new data are obtained for this study (i.e., updated search results), this section should be re-run by changing `echo=FALSE` to `echo=TRUE` in the knitr markdown argument.

```
# Clear workspace
rm(list=ls())

# Read BWR functions for Mac OS
source("/Users/beperron/Git/BibWrangleR/functions/piWrangleR.R")
source("/Users/beperron/Git/BibWrangleR/functions/packages.R")
# Set the path where original raw data are stored
setwd("/Users/beperron/Git/SocialWorkResearch")

# Set the working directory to store files created by BWR functions
my.path <- "/Users/beperron/Git/SocialWorkResearch"

# Wrangle the data with the BWR function suite
#piBWR.f(csv=FALSE, path=my.path)
#save(pi.df, file = "piArticles.R")
```

Initialize Windows workspace and functions for data wrangling

Initialize workspace and functions for analaysis

All the analyses performed involve the data that have been processed with the `BibWrangleR` functions. This section reads the processed data, loads the required packages, and does a quick quality check to ensure that

the same number of articles (i.e., records) contained in the original search match the number of articles in the transformed data.

```
rm(list=ls())
setwd("/Users/beperron/Git/SocialWorkResearch")
source("/Users/beperron/Git/BibWrangleR/functions/ggsurv.R")
load("piArticles.R")
library(dplyr)
library(ggplot2)
library(gridExtra)
library(survival)
library(grid)
library(png)


# Inspect dimensions of the data file (Rows X Columns)
dim(pi.df)
```

```
[1] 495415      3
```

```
# Inspect variable names of the data file
names(pi.df)
```

```
[1] "attributes" "articleID"  "record"
```

```
# How many unique article titles?  Ebsco Results of most current search is $n=24,314$. Do not proceed w
length(which(pi.df$attributes == "TI"))
```

```
[1] 24314
```

What is the overall number and names of journal titles?

```
unique.titles <- filter(pi.df, attributes == "SO")

# Number of unique titles
length(unique(unique.titles$record))
```

```
## [1] 89
```

```
# Unique titles
unique(unique.titles$record)
```

```
##  [1] "Journal of Ethnic & Cultural Diversity in Social Work: Innovation in Theory, Research & Practi
##  [2] "Journal of Sociology and Social Welfare"
##  [3] "Social Work & Christianity"
##  [4] "Journal of Gerontological Social Work"
##  [5] "Research on Social Work Practice"
##  [6] "Child & Family Social Work"
##  [7] "Australian Social Work"
```

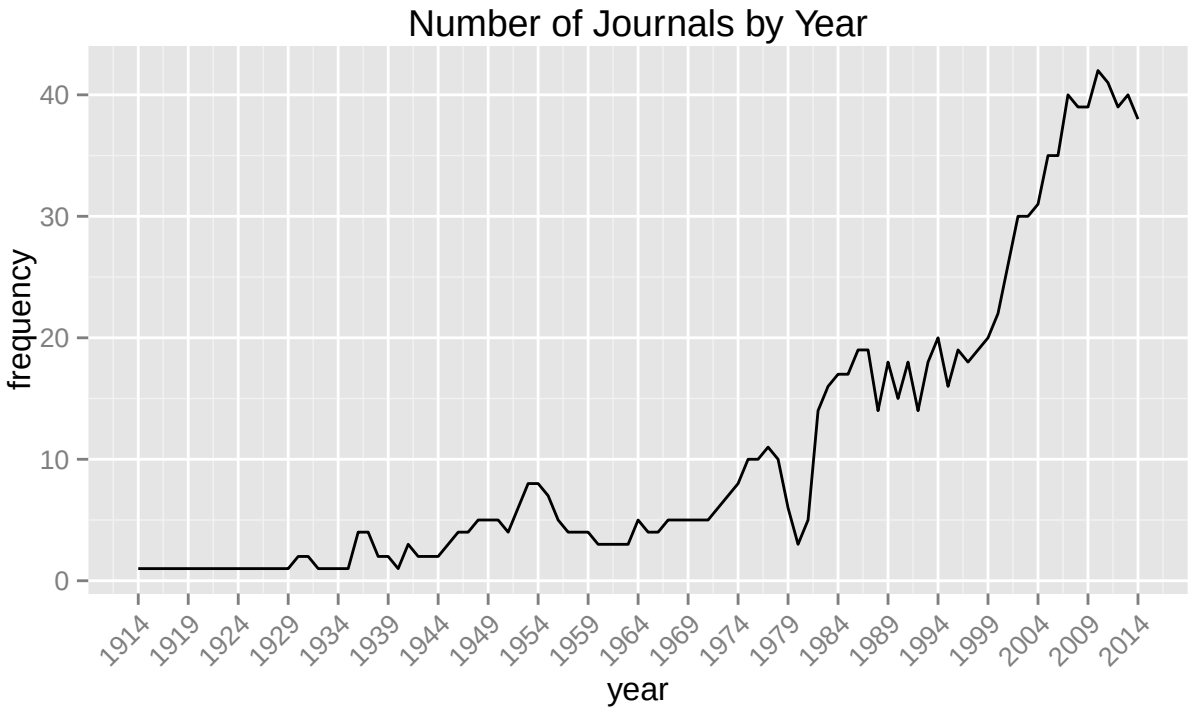
```
##  [8] "Social Work with Groups: A Journal of Community and Clinical Practice"
##  [9] "Practice: Social Work in Action"
## [10] "Journal of Gay & Lesbian Social Services: The Quarterly Journal of Community & Clinical Practi
## [11] "Smith College Studies in Social Work"
## [12] "Journal of Social Work Practice"
## [13] "Social Work in Health Care"
## [14] "Journal of Social Work Education"
## [15] "Children & Schools"
## [16] "Social Work"
## [17] "Child & Adolescent Social Work Journal"
## [18] "Clinical Social Work Journal"
## [19] "International Social Work"
## [20] "Journal of Social Work"
## [21] "Social Work Research"
## [22] "Social Work Education"
## [23] "Journal of Evidence-Based Social Work"
## [24] "Health & Social Work"
## [25] "Affilia: Journal of Women & Social Work"
## [26] "Qualitative Social Work: Research and Practice"
## [27] "Families in Society"
## [28] "Social Work in Mental Health"
## [29] "Ethics and Social Welfare"
## [30] "Journal of Religion & Spirituality in Social Work: Social Thought"
## [31] "Journal of HIV/AIDS & Social Services"
## [32] "Journal of Social Work Practice in the Addictions"
## [33] "British Journal of Social Work"
## [34] "School Social Work Journal"
## [35] "Journal of the Society for Social Work and Research"
## [36] "Journal of Social Work in End-of-Life & Palliative Care"
## [37] "International Journal of Social Welfare"
## [38] "Psychoanalytic Social Work"
## [39] "Administration in Social Work"
## [40] "The Journal of Baccalaureate Social Work"
## [41] "The Scientific Review of Mental Health Practice: Objective Investigations of Controversial and
## [42] "Social Work and Social Sciences Review"
## [43] "Journal of Gay & Lesbian Social Services: Issues in Practice, Policy & Research"
## [44] "Practice"
## [45] "Journal of Educational & Psychological Consultation"
## [46] "Rural Social Work"
## [47] "Journal of Technology in Human Services"
## [48] "Journal of Social Service Research"
## [49] "Journal of Applied Social Sciences"
## [50] "Early Child Development and Care"
## [51] "Computers in Human Services"
## [52] "The Clinical Supervisor"
## [53] "Children and Youth Services Review"
## [54] "Journal of Social Work Research and Evaluation"
## [55] "General Hospital Psychiatry"
## [56] "Canadian Journal on Aging"
## [57] "Social Casework"
## [58] "Journal of Multicultural Social Work"
## [59] "Journal of Analytic Social Work"
## [60] "Maatskaplike Werk/Social Work"
## [61] "Issues in Social Work Education"
```

```
## [62] "Journal of Teaching in Social Work"
## [63] "Social Work Research & Abstracts"
## [64] "Journal of Social Work & Human Sexuality"
## [65] "Journal of Independent Social Work"
## [66] "Employee Assistance Quarterly"
## [67] "Behavior Modification"
## [68] "Indian Journal of Social Work"
## [69] "Indian Journal of Psychiatric Social Work"
## [70] "British Journal of Psychiatric Social Work"
## [71] "Social Work in Education"
## [72] "Pediatric Social Work"
## [73] "Journal of Social Welfare"
## [74] "School Social Work Quarterly"
## [75] "Social Work Today"
## [76] "Journal of Psychiatric Social Work"
## [77] "Medical Social Work"
## [78] "Jewish Social Services Quarterly"
## [79] "Proceedings of the National Conference of Social Work"
## [80] "Journal of Social Casework"
## [81] "Social Work Yearbook"
## [82] "Social Work Technique"
## [83] "Journal of Social Work Process"
## [84] "Pennsylvania Social Work"
## [85] "International Conference of Social Work"
## [86] "Eugenics & Social Welfare Bull."
## [87] "New York State Department of Social Welfare, Division Publication"
## [88] "University of Washington Publications: Social Services"
## [89] "Eugenics and Social Welfare Bulletin"
```

Number of unique journal titles by year

```
journals.year <- tbl_df(pi.df)

year <- journals.year %>%
        filter(attributes == "YR") %>%
        select(id = articleID, year = record)

journals <- journals.year %>%
        filter(attributes == "SO") %>%
        select(id = articleID, journal.title = record)

n.journals.year <- journals %>%
        left_join(year) %>%
        group_by(year) %>%
        distinct(journal.title) %>%
        summarise(n = n())

journal.count <- ggplot(n.journals.year, aes(as.numeric(year), y=n, group=1)) +
    geom_line(colour="black") +
    #geom_point(colour="red") +
    theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
    xlab("year") +
```

```
    ylab("frequency") +
    ggtitle("Number of Journals by Year") +
    scale_x_continuous(breaks=seq(1914, 2014, 5))

journal.count
```



What journals published the most number of articles

```
n.so.yr <- filter(pi.df, attributes == "SO" | attributes == "YR")

n.so <- filter(pi.df, attributes == "SO") %>% mutate(title = record) %>%
        select(-attributes, -record)

n.yr <- filter(pi.df, attributes == "YR") %>% mutate(year = record ) %>%
        select(-attributes, -record)

n.so.yr <- left_join(n.so, n.yr) %>%
    group_by(title) %>%
    summarise(first = min(year), last = max(year), n.to.date = n()) %>%
    arrange(desc(n.to.date))
```

```
## Joining by: "articleID"
```

```
# 10 highest number of publications
head(n.so.yr, 10)
```

```
## Source: local data frame [10 x 4]
```

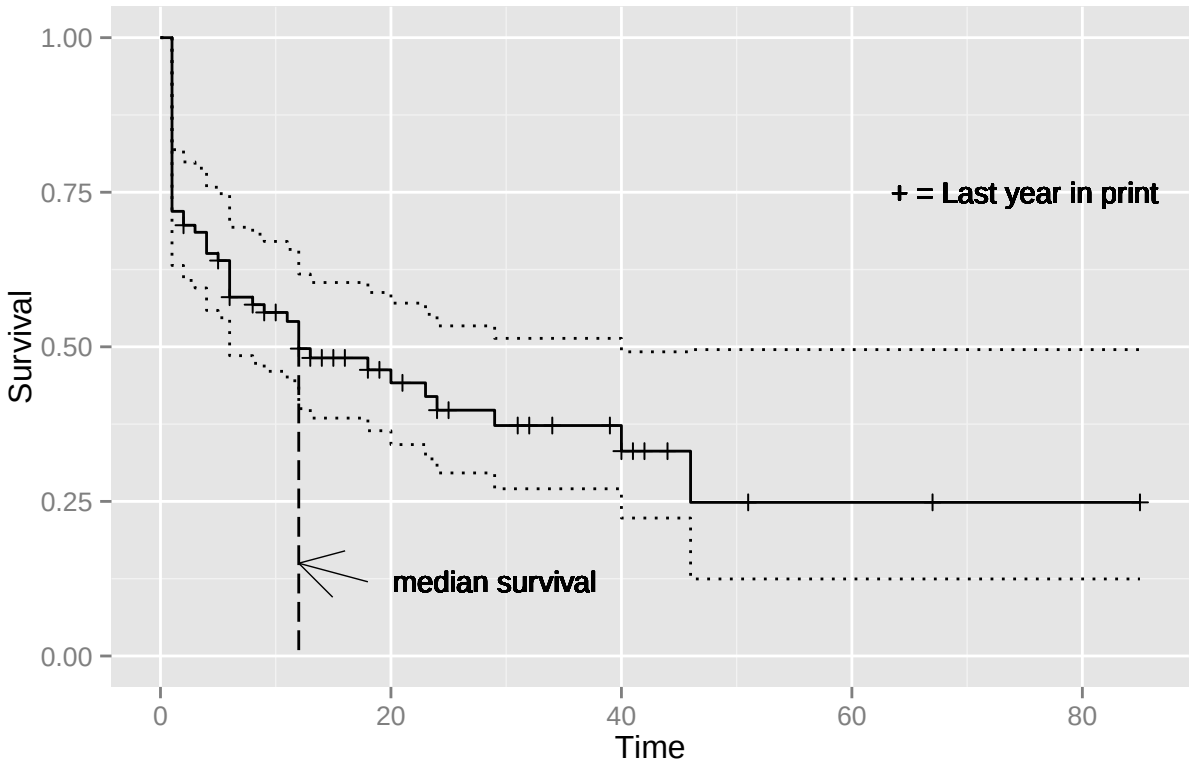
```
##
##                                 title first last n.to.date
## 1                          Social Work  1948 2014      1866
## 2        British Journal of Social Work  1971 2014      1456
## 3                  Families in Society  1990 2014      1211
## 4  Journal of Gerontological Social Work  1981 2014      1188
## 5            Social Work in Health Care  1975 2014      1171
## 6                       Social Casework  1950 1989      1095
## 7   Smith College Studies in Social Work  1930 2014      1075
## 8          Clinical Social Work Journal  1973 2014      1068
## 9      Research on Social Work Practice  1991 2014       986
## 10                 Health & Social Work  1976 2014       901
```

What is the lifespan of journals?

```
#10 longest running journals
longest.running <- n.so.yr %>%
       mutate(last = as.numeric(last), first = as.numeric(first),
              year.diff = last - first) %>%
       arrange(desc(year.diff)) %>%
       select(title, first, last, year.diff) %>%
       mutate(stop = year.diff, event = ifelse(as.numeric(last) != 2014, 1, 0)) %>%
       select(title, stop, event, as.numeric(first))

survival.journals <- survfit(Surv(longest.running$stop+1, longest.running$event) ~ 1)
median.survival <- data.frame(time = c(12,12), quant = c(.5,0))

head(longest.running)
```

```
## Source: local data frame [6 x 4]
##
##                                title stop event first
## 1 Smith College Studies in Social Work   84     0  1930
## 2                         Social Work   66     0  1948
## 3               Journal of Social Work   50     0  1964
## 4        Indian Journal of Social Work   45     1  1941
## 5       British Journal of Social Work   43     0  1971
## 6         Clinical Social Work Journal   41     0  1973
```

```
ggsurv(survival.journals) +
    geom_line(data = median.survival, aes(time, quant), linetype="longdash") +
    annotate("segment", x = 18, xend = 12, y = .12, yend = .15, size = .25, arrow =arrow()) +
    geom_text(x = 29, y =.12, label = "median survival", size = 4) +
    geom_text(x = 75, y =.75, label = "+ = Last year in print", size = 4)
```



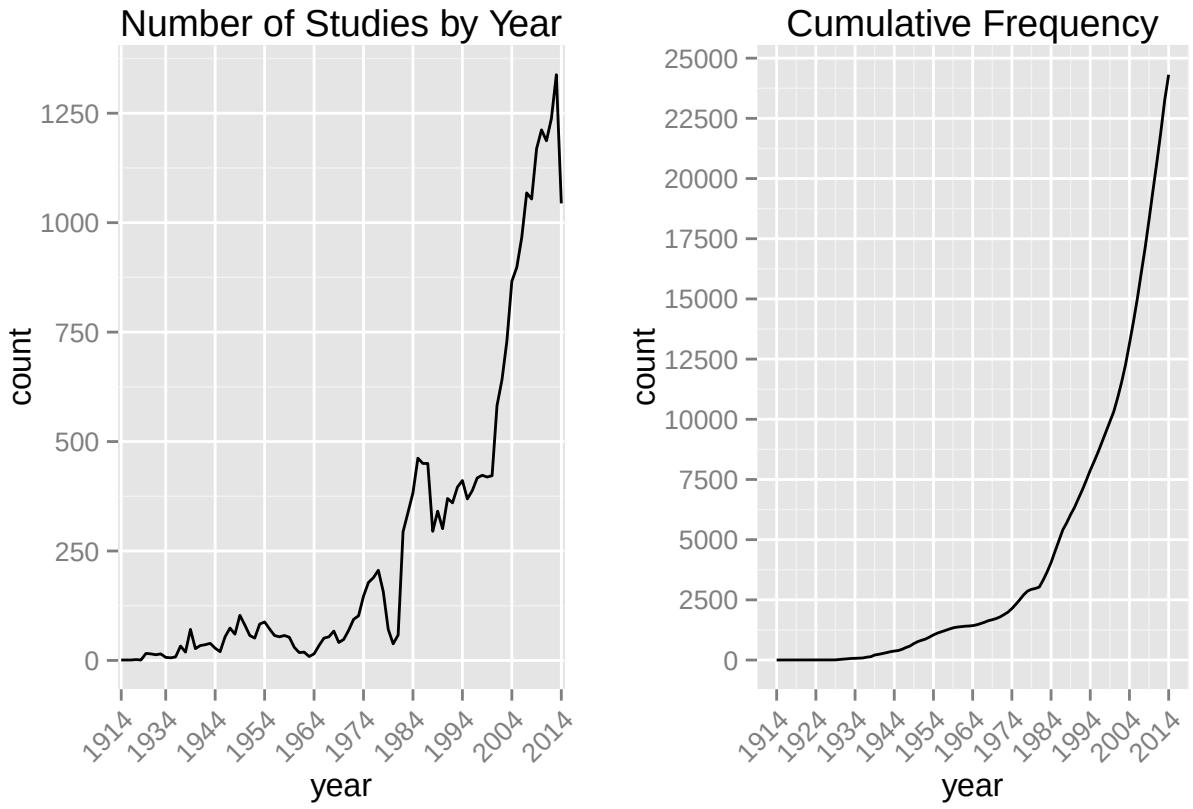
What is the number of articles published per year

```
n.articles.year <- filter(pi.df, attributes == "YR")
year.split <- split(n.articles.year, n.articles.year$record)
year.count <- unlist(lapply(year.split, nrow))
year.count <- year.count[order(names(year.count))]
years <- names(year.count)

df <- data.frame(years, year.count)
rownames(df) <- NULL

plot.article.count <- ggplot(df, aes(as.factor(years),
                     y = year.count, group=1)) +
    geom_line(colour="black") +
    #geom_point(colour="red") +
    theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
    xlab("year") +
    ylab("count") +
    ggtitle("Number of Studies by Year") +
    scale_x_discrete(breaks=c(seq(1914, 2014, 10))) +
    scale_y_continuous(breaks = c(seq(0, 2000, 250)))

df$years <- as.numeric(as.character(df$years))

plot.article.cumulative <- ggplot(df, aes(x = years, y = cumsum(year.count))) +
    geom_line() +
    theme(axis.text.x = element_text(angle=45, hjust=1)) +
```

```
    scale_x_continuous(breaks=pretty(df$years)) +
    xlab("year") +
    ylab("count") +
    scale_x_continuous(breaks = c(seq(1914,2014,10))) +
    scale_y_continuous(breaks = c(seq(0, 25000, 2500))) +
    ggtitle("Cumulative Frequency")
```

```
## Scale for 'x' is already present. Adding another scale for 'x', which will replace the existing scale
```

```
grid.arrange(plot.article.count, plot.article.cumulative, ncol=2)
```




```
# Print most recent ten years
head(df, 10)
```

```
##    years year.count
## 1   1914          1
## 2   1915          1
## 3   1918          1
## 4   1928          2
## 5   1929          1
## 6   1930         16
## 7   1931         15
## 8   1932         13
## 9   1933         15
## 10  1934          7
```

What are the topic areas (by Subject Terms)?

```
su.df <- filter(pi.df, attributes == "SU")

subject.terms <- stringr::str_split(su.df$record, pattern = ";")
subject.terms <- unlist(lapply(subject.terms, function(x) gsub(" ", "", x)))

subject.terms.total <- length(unlist(lapply(subject.terms,
                        function(x) gsub(" ", "", x))))

subject.terms.unique <- length(unique(subject.terms))

most.frequent <- as.data.frame(table(subject.terms))

most.frequent <- arrange(most.frequent, desc(Freq))

# Print 25 most commmon terms
head(most.frequent, 25)
```

```
                subject.terms Freq
1              SocialCasework 5794
2               SocialWorkers 2933
3         SocialWorkEducation 1696
4              SocialServices 1139
5                ChildWelfare  811
6               SocialSupport  602
7           CommunityServices  572
8                      Family  572
9                  Caregivers  571
10                 ChildAbuse  571
11            MentalDisorders  500
12               HumanFemales  493
13                  DrugAbuse  488
14            FamilyRelations  478
15                      Aging  474
16                        HIV  471
17                FosterCare  470
18                    Blacks  445
19       MentalHealthServices  441
20         HealthCareServices  440
21                MentalHealth  438
22             CopingBehavior  403
23               Intervention  400
24         GroupPsychotherapy  393
25 PsychotherapeuticProcesses  390
```

What are the topic areas over time (by Subject terms)?

```
decade <- filter(pi.df, attributes == "YR") %>%
    mutate(year = as.numeric(record)) %>% select(-record, -attributes)
```

```
decade$year <- cut(decade$year, breaks = 10, labels = c(1:10))

keywords <- pi.df %>%
        filter(attributes == "SU") %>%
        select(articleID = articleID, keywords = record)

keywords.decade <- keywords %>%
         left_join(decade)

library(plyr)
keywords.data.split <- dlply(keywords.decade, .(year))
detach(package:plyr)

terms.f <- function(x){
    split.terms <- stringr::str_split(x[,"keywords"], pattern =";")
    clean.terms <- lapply(split.terms, function(x) gsub(" ", "", x))
    }

keywords.decade <- lapply(keywords.data.split, terms.f)
keywords.decade <- lapply(keywords.decade, unlist)

temp <- lapply(keywords.decade, function(x) data.frame(table(x)))
temp <- lapply(temp, function(x) arrange(x, desc(Freq)))
lapply(temp, function(x) head(x,10))
```

```
## $`3`
##               x Freq
## 1 ChildGuidance    3
## 2        Agency    2
##
## $`4`
##                 x Freq
## 1          Agency   20
## 2   ChildGuidance    7
## 3 SmallBusinesses    1
##
## $`5`
##                       x Freq
## 1                Agency    9
## 2  EmotionalDisturbances    4
## 3          ChildGuidance    3
## 4                Clients    3
## 5        FamilyRelations    2
## 6          FamilyTherapy    2
## 7             Infidelity    2
## 8          SocialWorkers    2
## 9     AntisocialBehavior    1
## 10            ChildAbuse    1
##
## $`6`
##                       x Freq
## 1        SocialCasework  244
## 2     CommunityServices   45
```

```
## 3         SocialWorkers   45
## 4       FamilyRelations   41
## 5         FamilyTherapy   40
## 6             Treatment   36
## 7   PsychiatricPatients   30
## 8    GroupPsychotherapy   25
## 9                Family   23
## 10 ParentChildRelations   23
##
## $`7`
##                             x Freq
## 1              SocialCasework  527
## 2               SocialWorkers  228
## 3             GroupCounseling   92
## 4                HumanFemales   79
## 5               FamilyTherapy   72
## 6             FamilyRelations   71
## 7         SocialWorkEducation   67
## 8  PsychotherapeuticProcesses   65
## 9                     Parents   63
## 10         GroupPsychotherapy   62
##
## $`8`
##                     x Freq
## 1      SocialCasework  977
## 2       SocialWorkers  377
## 3      SocialServices  194
## 4     GroupCounseling  160
## 5       SocialSupport  148
## 6          ChildAbuse  140
## 7     FamilyRelations  135
## 8  GroupPsychotherapy  119
## 9       FamilyTherapy  118
## 10    MentalDisorders  111
##
## $`9`
##                      x Freq
## 1       SocialCasework 1118
## 2        SocialWorkers  550
## 3  SocialWorkEducation  279
## 4       SocialServices  277
## 5         ChildWelfare  198
## 6           ChildAbuse  172
## 7        SocialSupport  163
## 8           Caregivers  159
## 9            DrugAbuse  145
## 10   CommunityServices  139
##
## $`10`
##                      x Freq
## 1       SocialCasework 2927
## 2        SocialWorkers 1731
## 3  SocialWorkEducation 1242
## 4       SocialServices  620
```

```
## 5         ChildWelfare  573
## 6                Aging  369
## 7         Intervention  363
## 8               Family  343
## 9                  HIV  321
## 10          Caregivers  317
```

What are the most frequent topic areas (by author specified keywords)?

```
kp.df <- filter(pi.df, attributes == "KP")

subject.terms <- stringr::str_split(kp.df$record, pattern = ";")
subject.terms <- unlist(lapply(subject.terms, function(x) gsub(" ", "", x)))
subject.terms.total <- length(unlist(lapply(subject.terms,
                        function(x) gsub(" ", "", x))))

subject.terms.unique <- length(unique(subject.terms))

subject.terms.l <- list(subject.terms.total = subject.terms.total,
                      subject.terms.unique = subject.terms.unique)

most.frequent <- as.data.frame(table(subject.terms))

most.frequent <- arrange(most.frequent, desc(Freq))

# Print summary statistics
print(subject.terms.l)
```

```
$subject.terms.total
[1] 102493

$subject.terms.unique
[1] 46899
```

```
# Print 25 most frequent
head(most.frequent, 25)
```

```
             subject.terms Freq
1            socialworkers 1766
2               socialwork 1757
3      socialworkeducation  756
4       socialworkpractice  538
5           socialservices  340
6       socialworkstudents  314
7             mentalhealth  304
8                 children  283
9             childwelfare  255
10 CHILDHOODANDADOLESCENCE  236
11                    HIV  231
```

```
12           socialsupport  226
13             riskfactors  200
14           spirituality  199
15         decisionmaking  188
16             fostercare  187
17                  aging  179
18        domesticviolence  176
19           socialjustice  176
20              TECHNIQUES  175
21            intervention  174
22             adolescents  173
23             METHODOLOGY  173
24              SOCIALWORK  173
25           CHILDGUIDANCE  167
```

Most Frequent Author Keywords

```
decade <- filter(pi.df, attributes == "YR") %>%
    mutate(year = as.numeric(record)) %>% select(-record, -attributes)

decade$year <- cut(decade$year, breaks = 10, labels = c(1:10))

keywords <- pi.df %>%
        filter(attributes == "KP") %>%
        select(articleID = articleID, keywords = record)

keywords.decade <- keywords %>%
         left_join(decade)

library(plyr)
keywords.data.split <- dlply(keywords.decade, .(year))
detach(package:plyr)

terms.f <- function(x){
    split.terms <- stringr::str_split(x[,"keywords"], pattern =";")
    clean.terms <- lapply(split.terms, function(x) gsub(" ", "", x))
    }

keywords.decade <- lapply(keywords.data.split, terms.f)
keywords.decade <- lapply(keywords.decade, unlist)

temp <- lapply(keywords.decade, function(x) data.frame(table(x)))
temp <- lapply(temp, function(x) arrange(x, desc(Freq)))
lapply(temp, function(x) head(x,10))
```

```
## $`2`
##                                x Freq
## 1         CHILDHOODANDADOLESCENCE   38
## 2  SOCIALFUNCTIONSOFTHEINDIVIDUAL   27
## 3                           CHILD   19
## 4       NERVOUSANDMENTALDISORDERS   12
## 5                     DELINQUENCY    8
```

```
## 6                          FAMILY    8
## 7                  CHILDABILITIES    6
## 8  MOTHERATTITUDEANDBREASTFEEDING    6
## 9                     PERSONALITY    5
## 10                     ADJUSTMENT    4
##
## $`3`
##                                        x Freq
## 1                  CHILDHOODANDADOLESCENCE  155
## 2  GENERALSOCIALPROCESSES(INCL.ESTHETICS)   74
## 3                     FUNCTIONALDISORDERS   70
## 4                                   CHILD   63
## 5                                GUIDANCE   63
## 6                  CHILD(IV.MALADJUSTMENT   51
## 7                                 THERAPY)   51
## 8          CHILD(MALADJUSTMENTANDTHERAPY)   44
## 9                                    WORK   35
## 10                             ADJUSTMENT   32
##
## $`4`
##                   x Freq
## 1        SOCIALWORK  121
## 2     CHILDGUIDANCE  116
## 3    SOCIALCASEWORK  113
## 4        TECHNIQUES   95
## 5       METHODOLOGY   94
## 6  TREATMENTMETHODS   80
## 7              CASE   65
## 8        COUNSELING   62
## 9           SOCIAL   56
## 10         GUIDANCE   55
##
## $`5`
##                    x Freq
## 1         TECHNIQUES   80
## 2        METHODOLOGY   79
## 3      SOCIALWELFARE   72
## 4   TREATMENTMETHODS   42
## 5             FAMILY   39
## 6         SOCIALWORK   39
## 7      CHILDGUIDANCE   38
## 8      PSYCHOTHERAPY   34
## 9         COUNSELING   31
## 10 CRIME&DELINQUENCY   30
##
## $`6`
##                      x Freq
## 1                India   23
## 2        socialworkers   23
## 3           socialwork   13
## 4       socialcasework    9
## 5              clients    7
## 6       SOCIALCASEWORK    5
## 7           casereport    4
```

```
## 8             casework    4
## 9  COUNSELING&GUIDANCE    4
## 10        grouptherapy    4
##
## $`7`
##                               x Freq
## 1                 socialworkers   85
## 2                         India   83
## 3              literaturereview   50
## 4     implicationsforsocialwork   24
## 5                       children   22
## 6                           aged   21
## 7                        elderly   17
## 8                     socialwork   13
## 9             socialworkstudents   13
## 10 implicationsforsocialworkers   11
##
## $`8`
##                             x Freq
## 1              socialworkers  178
## 2     conferencepresentation  104
## 3  implicationsforsocialwork   75
## 4           literaturereview   62
## 5                     Israel   49
## 6                 casereport   44
## 7                    elderly   44
## 8                    England   44
## 9     socialworkimplications   41
## 10                     India   37
##
## $`9`
##                          x Freq
## 1           socialworkers  355
## 2              socialwork  317
## 3      socialworkpractice  119
## 4     socialworkeducation   88
## 5                 children   79
## 6            mentalhealth   62
## 7      socialworkstudents   58
## 8  conferencepresentation   55
## 9          socialservices   55
## 10                 Israel   52
##
## $`10`
##                      x Freq
## 1           socialwork 1411
## 2        socialworkers 1124
## 3  socialworkeducation  661
## 4   socialworkpractice  401
## 5       socialservices  283
## 6         mentalhealth  236
## 7   socialworkstudents  232
## 8         childwelfare  214
## 9                 HIV  210
```

```
## 10         riskfactors  175
```

Location of Studies

```
LO.df <- filter(pi.df, attributes == "LO")


subject.terms <- stringr::str_split(LO.df$record, pattern = ";")
subject.terms <- unlist(lapply(subject.terms, function(x) gsub(" ", "", x)))
subject.terms.total <- length(unlist(lapply(subject.terms, function(x) gsub(" ", "", x))))
subject.terms.unique <- length(unique(subject.terms))

subject.terms.l <- list(subject.terms.total = subject.terms.total,
                        subject.terms.unique = subject.terms.unique)

most.frequent <- as.data.frame(table(subject.terms))

location <- arrange(most.frequent, desc(Freq))

print(subject.terms.l)
```

```
$subject.terms.total
[1] 11076

$subject.terms.unique
[1] 204
```

```
print(location)
```

```
                    subject.terms Freq
1                              US 5308
2                   UnitedKingdom  696
3                       Australia  558
4                          Canada  542
5                          Israel  448
6                         England  388
7                           India  232
8                          Sweden  227
9                        HongKong  216
10                          China  158
11                     SouthAfrica  115
12                          Norway  101
13                      NewZealand   97
14                        Scotland   85
15                           Wales   84
16                         Ireland   83
17                         Germany   72
18                         Finland   70
19                     Netherlands   61
20                         Denmark   52
```

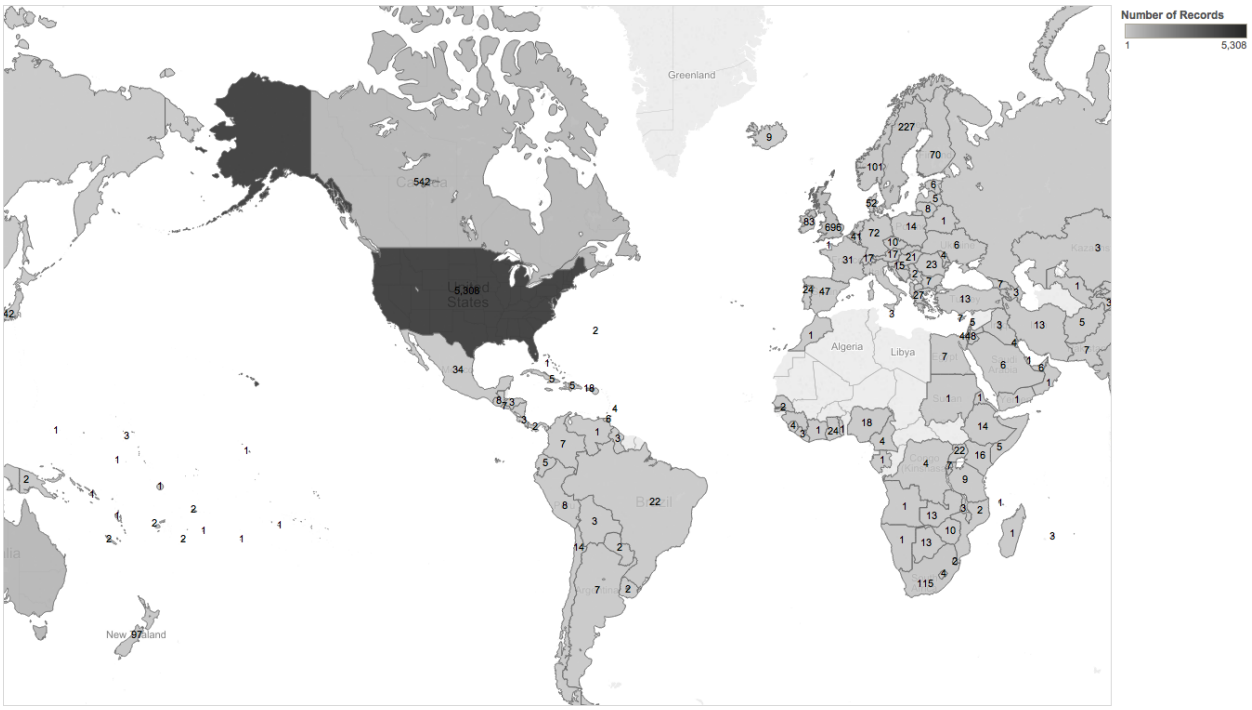
21	NorthernIreland	52
22	Spain	47
23	GreatBritain	45
24	Italy	44
25	Japan	42
26	Belgium	41
27	Singapore	41
28	Taiwan	41
29	Africa	38
30	Mexico	34
31	Europe	31
32	France	31
33	Greece	27
34	Russia	27
35	Ghana	24
36	Korea	24
37	Portugal	24
38	Romania	23
39	Brazil	22
40	Thailand	22
41	Uganda	22
42	Hungary	21
43	Asia	18
44	Nigeria	18
45	PuertoRico	18
46	SouthKorea	18
47	Vietnam	18
48	Austria	17
49	Switzerland	17
50	Kenya	16
51	Croatia	15
52	Chile	14
53	Ethiopia	14
54	NorthAmerica	14
55	Poland	14
56	Botswana	13
57	Iran	13
58	Turkey	13
59	Zambia	13
60	Malaysia	12
61	Slovenia	11
62	CzechRepublic	10
63	Zimbabwe	10
64	Bangladesh	9
65	Iceland	9
66	Tanzania	9
67	Guatemala	8
68	Lithuania	8
69	Luxembourg	8
70	Nepal	8
71	Peru	8
72	Philippines	8
73	Argentina	7
74	Bulgaria	7

75	Caribbean	7
76	Colombia	7
77	Cyprus	7
78	Egypt	7
79	ElSalvador	7
80	Georgia	7
81	Jordan	7
82	Pakistan	7
83	Palestine	7
84	Rwanda	7
85	USSR	7
86	Albania	6
87	Estonia	6
88	SaudiArabia	6
89	SriLanka	6
90	TrinidadandTobago	6
91	Ukraine	6
92	UnitedArabEmirates	6
93	Afghanistan	5
94	Cambodia	5
95	Cuba	5
96	DominicanRepublic	5
97	Ecuador	5
98	Indonesia	5
99	Latvia	5
100	Lebanon	5
101	Mongolia	5
102	Slovakia	5
103	Somalia	5
104	Barbados	4
105	Cameroon	4
106	DemocraticRepublicofCongo	4
107	Kuwait	4
108	Lesotho	4
109	Moldova	4
110	Nicaragua	4
111	Oceania/PacificIslands	4
112	SierraLeone	4
113	SouthAmerica	4
114	Azerbaijan	3
115	Bolivia	3
116	Bosnia-Herzegovina	3
117	CentralAmerica	3
118	CostaRica	3
119	Czechoslovakia	3
120	Guyana	3
121	Haiti	3
122	Honduras	3
123	Iraq	3
124	Jamaica	3
125	Kazakhstan	3
126	Liberia	3
127	Malawi	3
128	Malta	3

129	MarshallIslands	3
130	Mauritius	3
131	Tajikistan	3
132	Yugoslavia	3
133	Appalachia	2
134	Bermuda	2
135	Bhutan	2
136	EasternEurope	2
137	Fiji	2
138	Gambia	2
139	Kyrgyzstan	2
140	MiddleEast	2
141	Mozambique	2
142	Myanmar	2
143	NewCaledonia	2
144	Palau	2
145	Panama	2
146	PapuaNewGuinea	2
147	Paraguay	2
148	RepublicofSerbia	2
149	Samoa	2
150	Scandinavia	2
151	Swaziland	2
152	Tonga	2
153	Uruguay	2
154	WesternEurope	2
155	Angola	1
156	Armenia	1
157	Bahamas	1
158	Bahrain	1
159	BalticStates	1
160	Belarus	1
161	Brunei	1
162	Burundi	1
163	ChannelIslands	1
164	CommonwealthofIndependentStates	1
165	Comoros	1
166	CookIslands	1
167	Eritrea	1
168	FrenchPolynesia	1
169	Gabon	1
170	Grenada	1
171	Guinea	1
172	IvoryCoast	1
173	Kiribati	1
174	Laos	1
175	LatinAmerica	1
176	Liechtenstein	1
177	Macau	1
178	Macedonia	1
179	Madagascar	1
180	Maldives	1
181	Micronesia(FederatedStatesof)	1
182	Morocco	1

183	Namibia	1
184	Nauru	1
185	Niue	1
186	NorthKorea	1
187	Oman	1
188	Qatar	1
189	RepublicofCongo	1
190	Senegal	1
191	SerbiaandMontenegro	1
192	SlovakRepublic	1
193	SolomonIslands	1
194	StKitts	1
195	Sudan	1
196	Togo	1
197	Tuvalu	1
198	USVirginIslands	1
199	Uzbekistan	1
200	Vanuatu	1
201	Venezuela	1
202	WestBank	1
203	WestIndies	1
204	Yemen	1

```
img2 <- readPNG("/Users/beperron/Git/SocialWorkResearch/Chloro.png")
grid.raster(img2)
```



Location of studies over time

```
top.10.countries <- head(location, 10)
top.10.countries <- top.10.countries$subject.terms
top.10.countries <- levels(droplevels(top.10.countries))



year <- filter(pi.df, attributes == "YR") %>%
    mutate(year = as.numeric(record)) %>% select(-record, -attributes)


location <- pi.df %>%
        filter(attributes == "LO") %>%
        select(articleID = articleID, keywords = record) %>%
        filter(keywords %in% top.10.countries)


location.year <- location %>%
         left_join(year)



plot.article.cumulative <- ggplot(df, aes(x = years, y = cumsum(year.count))) +
    geom_line() +
    theme(axis.text.x = element_text(angle=45, hjust=1)) +
    scale_x_continuous(breaks=pretty(df$years)) +
    xlab("year") +
    ylab("count") +
    scale_x_continuous(breaks = c(seq(1914,2014,10))) +
    scale_y_continuous(breaks = c(seq(0, 25000, 2500))) +
    ggtitle("Cumulative Frequency")
```

Methodology

It is easy to explore some of the different fields within the PsychInfo data frame. For example, each record has one or more subject terms (from the article keywords). The total number, unique number, and most frequently occuring key words can be easily computed.

```
MD.df <- filter(pi.df, attributes == "MD")

subject.terms <- stringr::str_split(MD.df$record, pattern = ";")
subject.terms <- unlist(lapply(subject.terms, function(x) gsub(" ", "", x)))
subject.terms.total <- length(unlist(lapply(subject.terms, function(x) gsub(" ", "", x))))
subject.terms.unique <- length(unique(subject.terms))

subject.terms.l <- list(subject.terms.total = subject.terms.total,
                        subject.terms.unique = subject.terms.unique)

most.frequent <- as.data.frame(table(subject.terms))

most.frequent <- arrange(most.frequent, desc(Freq))
most.frequent.t <- head(most.frequent, 50)
```

```r
print(subject.terms.l)
```

```
$subject.terms.total
[1] 25380

$subject.terms.unique
[1] 21
```

```r
print(most.frequent.t)
```

```
                   subject.terms  Freq
1                  EmpiricalStudy 11741
2                QuantitativeStudy  4296
3                 QualitativeStudy  3455
4                        Interview  2300
5                 LiteratureReview   879
6                LongitudinalStudy   584
7                      FocusGroup   469
8                ClinicalCaseStudy   423
9             NonclinicalCaseStudy   319
10                 FollowupStudy   288
11                    FieldStudy   133
12               SystematicReview   109
13             RetrospectiveStudy   100
14 TreatmentOutcome/ClinicalTrial    84
15               ProspectiveStudy    69
16                   MetaAnalysis    63
17              MathematicalModel    34
18        ExperimentalReplication    27
19           ScientificSimulation     5
20                  BrainImaging     1
21                     TwinStudy     1
```

Methodology

```r
decade <- filter(pi.df, attributes == "YR") %>%
    mutate(year = as.numeric(record)) %>% select(-record, -attributes)

decade$year <- cut(decade$year, breaks = 20, labels = c(1:20))

keywords <- pi.df %>%
        filter(attributes == "MD") %>%
        select(articleID = articleID, keywords = record)

keywords.decade <- keywords %>%
         left_join(decade)
```

```
## Joining by: "articleID"
```

```
library(plyr)
```

```
## ------------------------------------------------------------------------------
## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)
## ------------------------------------------------------------------------------
##
## Attaching package: 'plyr'
##
## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
keywords.data.split <- dlply(keywords.decade, .(year))
detach(package:plyr)

terms.f <- function(x){
    split.terms <- stringr::str_split(x[,"keywords"], pattern =";")
    clean.terms <- lapply(split.terms, function(x) gsub(" ", "", x))
    }

keywords.decade <- lapply(keywords.data.split, terms.f)
keywords.decade <- lapply(keywords.decade, unlist)
lapply(keywords.decade, function(x) length(unique(x)))
```

```
## $`4`
## [1] 1
##
## $`5`
## [1] 1
##
## $`6`
## [1] 1
##
## $`7`
## [1] 1
##
## $`8`
## [1] 1
##
## $`9`
## [1] 1
##
## $`10`
## [1] 5
##
## $`11`
## [1] 6
##
## $`12`
```

```
## [1] 6
##
## $`13`
## [1] 5
##
## $`14`
## [1] 8
##
## $`15`
## [1] 10
##
## $`16`
## [1] 8
##
## $`17`
## [1] 16
##
## $`18`
## [1] 16
##
## $`19`
## [1] 18
##
## $`20`
## [1] 21
```

```
temp <- lapply(keywords.decade, function(x) data.frame(table(x)))
temp <- lapply(temp, function(x) arrange(x, desc(Freq)))
lapply(temp, function(x) head(x,10))
```

```
## $`4`
##           x Freq
## 1 Interview    1
##
## $`5`
##           x Freq
## 1 Interview    5
##
## $`6`
##           x Freq
## 1 Interview    3
##
## $`7`
##           x Freq
## 1 Interview    6
##
## $`8`
##           x Freq
## 1 Interview   10
##
## $`9`
##           x Freq
## 1 Interview    9
##
```

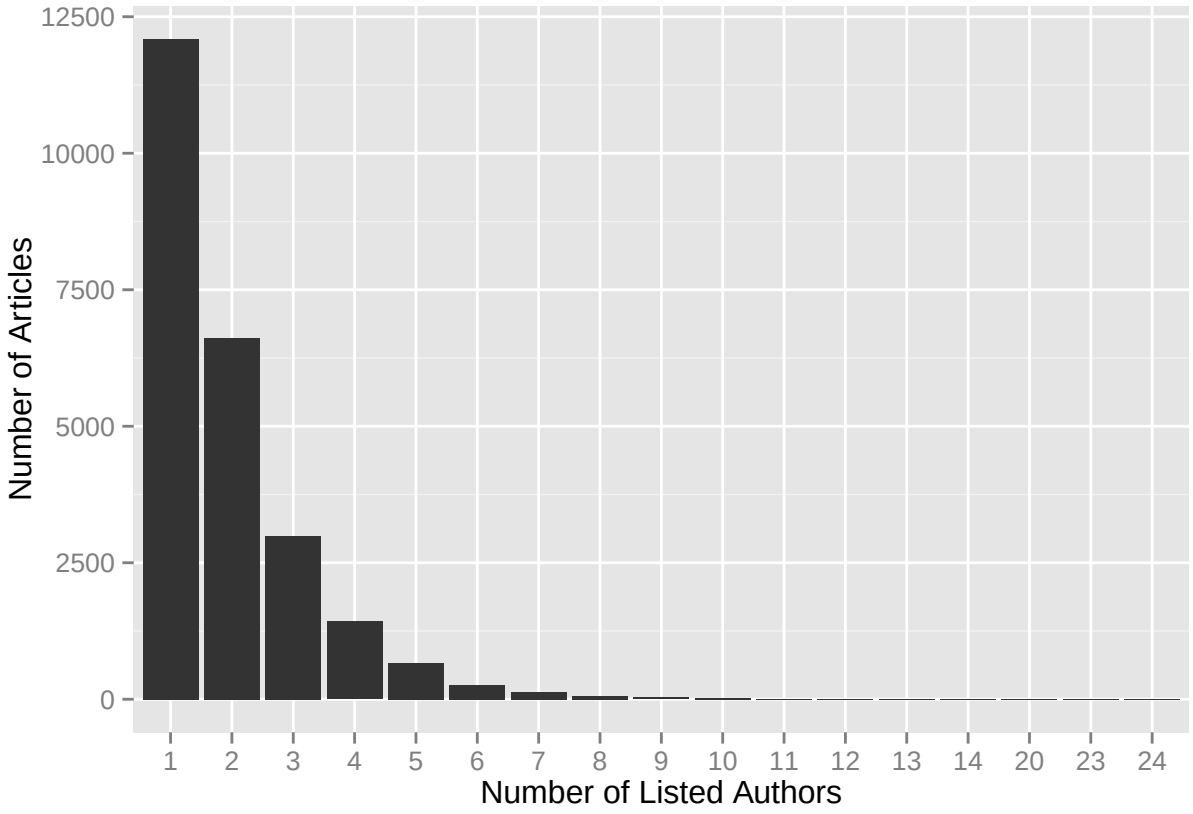
```
## $`10`
##                   x Freq
## 1    EmpiricalStudy    9
## 2         Interview    5
## 3 QuantitativeStudy    4
## 4  LiteratureReview    1
## 5  QualitativeStudy    1
##
## $`11`
##                   x Freq
## 1    EmpiricalStudy   20
## 2 QuantitativeStudy    8
## 3 ClinicalCaseStudy    6
## 4         Interview    5
## 5     FollowupStudy    3
## 6  LiteratureReview    2
##
## $`12`
##                   x Freq
## 1 ClinicalCaseStudy   10
## 2  LiteratureReview    8
## 3         Interview    5
## 4    EmpiricalStudy    2
## 5     FollowupStudy    1
## 6 LongitudinalStudy    1
##
## $`13`
##                   x Freq
## 1  LiteratureReview   19
## 2 ClinicalCaseStudy   13
## 3    EmpiricalStudy    7
## 4     FollowupStudy    7
## 5         Interview    2
##
## $`14`
##                   x Freq
## 1    EmpiricalStudy  535
## 2  LiteratureReview   49
## 3 ClinicalCaseStudy   22
## 4     FollowupStudy    6
## 5         Interview    6
## 6 LongitudinalStudy    1
## 7      MetaAnalysis    1
## 8  SystematicReview    1
##
## $`15`
##                                 x Freq
## 1                  EmpiricalStudy  960
## 2               ClinicalCaseStudy   47
## 3                LiteratureReview   30
## 4                   FollowupStudy   17
## 5                       Interview   10
## 6              LongitudinalStudy    7
## 7         ExperimentalReplication    4
```

```
## 8                    MetaAnalysis    3
## 9  TreatmentOutcome/ClinicalTrial    2
## 10           NonclinicalCaseStudy    1
##
## $`16`
##                          x Freq
## 1          EmpiricalStudy 1052
## 2        LiteratureReview   40
## 3           FollowupStudy   24
## 4       ClinicalCaseStudy   21
## 5       LongitudinalStudy   20
## 6               Interview    9
## 7           MetaAnalysis    3
## 8 ExperimentalReplication    1
##
## $`17`
##                                 x Freq
## 1                 EmpiricalStudy 1114
## 2               LiteratureReview   70
## 3              LongitudinalStudy   56
## 4              ClinicalCaseStudy   42
## 5                  FollowupStudy   29
## 6           NonclinicalCaseStudy   26
## 7  TreatmentOutcome/ClinicalTrial   15
## 8                      Interview   10
## 9         ExperimentalReplication    5
## 10              QualitativeStudy    5
##
## $`18`
##                                 x Freq
## 1                 EmpiricalStudy 1757
## 2              QuantitativeStudy  487
## 3               QualitativeStudy  372
## 4               LiteratureReview  184
## 5           NonclinicalCaseStudy   81
## 6              ClinicalCaseStudy   72
## 7              LongitudinalStudy   72
## 8                  FollowupStudy   40
## 9                      Interview   28
## 10 TreatmentOutcome/ClinicalTrial   20
##
## $`19`
##                        x Freq
## 1        EmpiricalStudy 2682
## 2     QuantitativeStudy 1671
## 3      QualitativeStudy 1220
## 4      LiteratureReview  176
## 5             Interview  158
## 6     LongitudinalStudy  154
## 7     ClinicalCaseStudy   89
## 8  NonclinicalCaseStudy   89
## 9         FollowupStudy   61
## 10           FocusGroup   30
##
```

```
## $`20`
##                         x Freq
## 1        EmpiricalStudy 3603
## 2     QuantitativeStudy 2122
## 3             Interview 2028
## 4      QualitativeStudy 1857
## 5            FocusGroup  428
## 6      LiteratureReview  300
## 7     LongitudinalStudy  273
## 8  NonclinicalCaseStudy  122
## 9             FieldStudy  116
## 10    ClinicalCaseStudy  101
```

Number of authors

```
n.authors.article <- pi.df %>%
    filter(attributes == "AU") %>%
    select(id = articleID, author= record) %>%
    mutate(id = as.numeric(id))

n_authors <- n.authors.article %>%
        group_by(id) %>%
        summarise(n = n())

ggplot(n_authors, aes(x = factor(n))) +
    geom_bar() +
    stat_bin(binwidth=1) +
    xlab("Number of Listed Authors") +
    ylab("Number of Articles")
```




```
summary(n_authors$n)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    1.00    2.00    1.94    2.00   24.00
```

Number of authors over time

This figure shows the average number of authors, along with the standard deviation as the ribbon around the average. Note that there is a possible problem in these data, with a single article listing a huge number. That can be corrected at a later time.

```
df.2 <- tbl_df(pi.df)
year <- df.2 %>%
        filter(attributes == "YR") %>%
        select(id = articleID, year = record)

authors <- df.2 %>%
        filter(attributes == "AU") %>%
        select(id = articleID, author = record)

n_authors <- authors %>%
        group_by(id) %>%
        summarise(n=n())

n_authors <- n_authors %>%
        left_join(year) %>%
```

```
        group_by(year) %>%
        summarise(median.n = median(n),
                  average.n = mean(n),
                  min.n = min(n),
                  max.n = max(n),
                  std.dev  = sd(n) )

plot.author.count2 <- ggplot(n_authors, aes(as.numeric(year), y=average.n, group=1)) +
    geom_line(colour="black") +
    geom_ribbon(aes(ymin = average.n-std.dev, ymax=average.n+std.dev), alpha=.2)

head(n_authors, 20)
```

```
Source: local data frame [20 x 6]

   year median.n average.n min.n max.n std.dev
1  1914        1     1.000     1     1      NA
2  1915        1     1.000     1     1      NA
3  1918        1     1.000     1     1      NA
4  1928        1     1.000     1     1  0.0000
5  1929        1     1.000     1     1      NA
6  1930        1     1.000     1     1  0.0000
7  1931        1     1.267     1     2  0.4577
8  1932        1     1.077     1     2  0.2774
9  1933        1     2.600     1    23  5.6543
10 1934        1     1.143     1     2  0.3780
11 1935        1     1.167     1     2  0.4082
12 1936        2     1.750     1     3  0.7071
13 1937        1     1.152     1     4  0.5658
14 1938        1     1.632     1     6  1.3000
15 1939        1     1.070     1     3  0.3082
16 1940        1     1.111     1     3  0.4237
17 1941        1     1.118     1     3  0.4093
18 1942        1     1.139     1     5  0.6825
19 1943        1     1.077     1     4  0.4804
20 1944        1     1.000     1     1  0.0000
```

```
tail(n_authors, 20)
```

```
Source: local data frame [20 x 6]

   year median.n average.n min.n max.n std.dev
1  1995        2     1.921     1     8   1.214
2  1996        2     1.879     1     9   1.185
3  1997        1     1.739     1     6   1.038
4  1998        2     1.934     1     9   1.271
5  1999        1     1.800     1     9   1.185
6  2000        2     1.853     1     8   1.138
7  2001        1     1.842     1     8   1.179
8  2002        2     1.966     1     7   1.155
9  2003        2     1.936     1    20   1.371
10 2004        2     1.928     1     8   1.201
11 2005        2     2.053     1    11   1.388
```

```
12 2006        2     2.122     1    12   1.402
13 2007        2     2.138     1    14   1.488
14 2008        2     2.145     1    12   1.429
15 2009        2     2.246     1    12   1.486
16 2010        2     2.241     1    12   1.488
17 2011        2     2.340     1    13   1.555
18 2012        2     2.305     1    24   1.583
19 2013        2     2.459     1    14   1.660
20 2014        2     2.430     1    12   1.618
```

```
plot.author.count2
```

How Many International Contributors?

This section shows a proof of concept – that is, we can potentially extract all the countries from the author affiliation `AF` tag in the data set. This involves using a set of regular expressions for the extraction. Here I have hard-coded a few countries, but I can obtain a file of all countries and use that to automate the process. We will need to look at the raw data to ensure that the author affiliations have remained in a consistent format throughtout the entirety of the study.

```
df.affiliations <- pi.df %>%
                filter(attributes == "AF")

us.aff <- ifelse(grepl("US", df.affiliations$record, perl=TRUE) == TRUE, "US",
          ifelse(grepl("Canada", df.affiliations$record, perl=TRUE ) == TRUE, "Canada",
          ifelse(grepl("Kong", df.affiliations$record, perl=TRUE ) == TRUE, "Hong Kong",
          ifelse(grepl("China", df.affiliations$record, perl=TRUE) == TRUE, "China",
          ifelse(grepl("Israel", df.affiliations$record, perl=TRUE) == TRUE, "Israel", "Other" )))))
```

```
affiliations <- data.frame(cbind(df.affiliations,us.aff))

ggplot(data=df.affiliations, aes(x = factor(us.aff))) + geom_bar()
```