Supplementary Information

Ye Zhang

Biomedical Engineering Department, Johns Hopkins University, Baltimore, MD 21218

Preliminary Exploration

After tidying up the downloaded 500 abstracts, I summarized all the single word, "bigram" (knowns as two words combination) and "trigram" (known as three words combination) respectively and calculated their appearance frequency in these 500 abstracts. The top 20 most frequently mentioned "word", "bigram" and "trigram" were shown in **Supplementary Table 1**, **2** and **3**.

Table 1: Top 20 single word in sample abstracts

word	n
study	567
patients	521
data	460
analysis	439
risk	354
significant	343
ci	333
model	326
health	298
statistical	295
results	293
based	270
studies	262
age	244
compared	233
statistically	229
disease	224
treatment	221
time	218
models	216

Table 2: Top 20 bigram in sample abstracts

bigram	n
statistically significant	159
risk factors	58
meta analysis	56
breast cancer	44
logistic regression	44
mental health	41
significant differences	37
statistical analysis	37
aor ci	36
public health	35
confidence interval	34

bigram	n
cross sectional	33
gene expression	32
odds ratio	32
significant difference	32
statistical significance	30
controlled trials	25
lung cancer	25
randomized controlled	25
results suggest	25

Table 3: Top 20 trigram in sample abstracts

trigram	n
body mass index	19
statistically significant differences	19
statistically significant difference	16
confidence interval ci	15
logistic regression analysis	15
cross sectional study	12
randomized controlled trials	11
acetate pet mri	10
children aged months	10
clif sofa score	10
genome wide association	10
antenatal magnesium sulphate	9
autism spectrum disorder	9
common mental disorder	9
mass index bmi	9
negative breast cancer	9
randomized controlled trial	9
swift wound app	9
born completed weeks	8
breast cancer risk	8

Results

The number of articles reporting a statistical method in each research fields was counted in the **Supplementary Table 4**.

Table 4: Number of articles reporting statistical methods in each field

methods	General	Biology	Medicine	Computational Biology	Genetics	Neglected tropical diseases	Pathogen
logistic regression	6189	17	97	57	80	295	24
meta analysis	2608	23	58	47	126	83	41
bootstrap	3230	40	32	91	120	185	71
ANOVA	14283	99	24	60	278	282	376
clustering	8145	102	89	352	357	373	184
bayesian	2328	30	24	154	66	137	51
t-test	13592	31	23	79	445	297	434
linear regression	9137	79	122	244	250	306	138
machine learning	868	10	5	81	22	29	16
maximum likelihood	4823	79	40	270	191	270	101
neural network	704	31	0	184	12	7	4
random forest	3389	14	27	64	35	124	32
support vector machine	591	10	3	68	18	27	8
MCMC	895	8	4	53	22	71	22