SAS® GLOBAL FORUM 2016

IMAGINE. CREATE. INNOVATE.

A SAS® Program to Analyze High Dimensional Sparse Counts Data Using Lumping and Splitting Approach

#SASGF

A SAS® Program to Analyze High Dimensional Sparse Counts Data Using Lumping and Splitting Approach

Xiaoli Lu

VA Cooperative Studies Program, Perry Point, MD 21902

ABSTRACT

It is common for hundreds or even thousands of clinical endpoints to be collected from individual subjects, but events for the majority of clinical endpoints are rare. The challenge to analyzing high dimensional sparse data is how to balance analytical consideration for statistical inference and clinical interpretation with precise meaningful outcome of interest at intra-categorical and inter-categorical levels. Lumping or grouping similar rare events into a composite category has the statistical advantage of increasing testing power, reducing multiplicity size, and avoiding competing risk problem; however, too much or inappropriate lumping could jeopardize the clinical interpretation of outcomes and external validity. Whereas splitting or keeping each individual event at its basic clinical meaningful category can overcome the drawbacks of lumping, this practice may create analytical issues of increasing type II error, multiplicity size, competing risks, and having a large proportion of endpoints with rare events. It seems that lumping and splitting are diametrically opposed approaches, but in fact, they are complementary. Both are essential for high dimensional data analysis. This paper describes the steps required for the lumping and splitting analysis, and presents SAS code that can be used to implement each step.

BACKGROUND

In clinical studies, there could be multiple efficacy/safety endpoints, such as imaging data, genomic data, cognitive function data, adverse event data etc. To analyze these kinds of multiple count endpoint data, one often has to deal with two issues: data high dimensionality and sparsity.

Multiple Sparse Counts Data

- Non-negative integers
- Represent the number of occurrences within a fixed time
- Both events and duration can be parameterized
- Variance increases with the expected number of events

Data Capture

ID	Onset Time	End Time	Event	Other Variables … …
S_1	mmddyyyy	mmddyyyy	E_7	… …
S_2	mmddyyyy	mmddyyyy	E_5	… …
…	…	…	…	… …
S_{n-1}	mmddyyyy	mmddyyyy	E_8	… …
S_n	mmddyyyy	mmddyyyy	E_1	… …

m Events

n Subjects

ID	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	…	E_{m-1}	E_m
S_1	0	0	0	0	0	0	1	0	…	0	0
S_2	0	1	0	0	0	0	0	0	…	0	0
S_3	0	0	0	0	0	0	0	0	…	0	0
S_4	0	0	0	0	0	0	0	0	…	0	0
S_5	0	0	0	0	0	0	0	0	…	0	0
S_6	0	0	0	0	0	0	0	0	…	0	0
S_7	0	0	0	0	0	0	0	0	…	0	0
S_8	0	0	0	0	0	0	0	0	…	2	0
…	…	…	…	…	…	…	…	…	…	…	…
S_{n-1}	0	0	0	0	0	0	0	1	…	0	0
S_n	1	0	0	0	0	0	0	0	…	0	0

- n subjects, m endpoints, m >>n
- too many zeros

Data Formation

* Multiple Counts Data * Longitudinal Data * Time-to-Event Data

Statistical Issues

Power Issue (Rare Incidence) Multiplicity Issue (Many Features) Competing Risk Issue (Multiple Events)

METHODS

Data Process

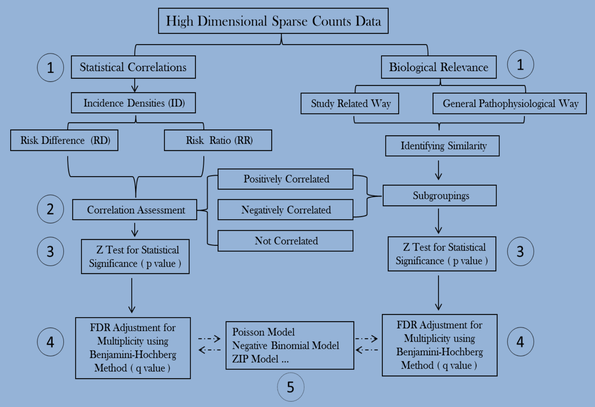
- Code events according to clinical and biological relevance either based on study related symptoms or pathophysiological pathway, then identify the similar elements to form lumping groups
- Cluster events based on statistical correlations by computing event incidence densities and further estimate risk differences or risk ratios
- Assess correlations between lumped events and intervention groups
- Stratify the lumped events based on positive or negative correlations

Analysis Approach

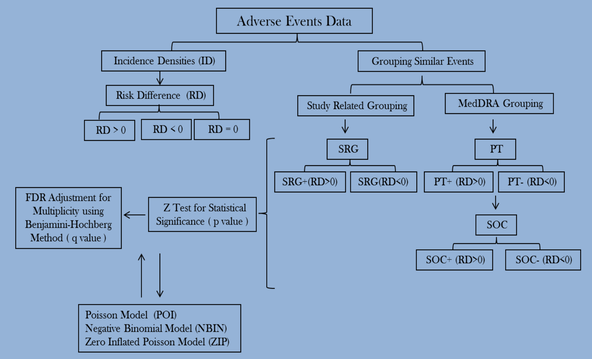
- Generate raw p-values using proportion Z tests or exact tests depending on whether number of events are still rare after lump up
- Generate q-values by adjusting false discovery rate using Benjamini-Hochberg method to control inflated type I error caused by multiplicity
- Validate the results using Poisson model or negative binomial model

Example

- A placebo-controlled, double-blind, randomized multi-center trial
- n = 212
- m = 864 adverse events as multiple endpoints from 1348 reported events



Data Process and Analysis Strategy



Example Analysis Schema

A SAS® Program to Analyze High Dimensional Sparse Counts Data Using Lumping and Splitting Approach

Xiaoli Lu

VA Cooperative Studies Program, Perry Point, MD 21902

RESULTS

1. Analysis based on individual events (Splitting Approach)

#	Individual Events (IE)	ID (trt)	ID (plb)	RD	Z	p	q
1	AUTONOMIC DYSREFLEXIA	0.1167	0.0042	0.1124	7.7076	0.0000	0.0000
2	ELEVATED TEMPERTURE	0.0316	0.0021	0.0295	3.6588	0.0003	NS
3	ELEVATED ALT	0.0257	0.0000	0.0257	3.6528	0.0003	NS
4	ANAEMIA	0.0178	0.0000	0.0178	3.0271	0.0025	NS
5	HYPERGLYCEMIA	0.0099	0.0000	0.0099	2.2472	0.0246	NS
6	HYPONATREMIA	0.0099	0.0000	0.0099	2.2472	0.0246	NS
7	HYPERKALAEMIA	0.0138	0.0021	0.0117	2.0885	0.0368	NS
8	FEVER	0.0178	0.0042	0.0136	2.0539	0.0400	NS
9	NAUSEA AND VOMITING	0.0000	0.0085	-0.0085	2.0086	0.0446	NS
10	ELEVATED AST	0.0079	0.0000	0.0079	2.0080	0.0447	NS
11	ELEVATED PLATELETS	0.0079	0.0000	0.0079	2.0080	0.0447	NS

2. Analysis based on grouped events (Lumping Approach)

#	Preferred Terms (PT)	ID (trt)	ID (plb)	RD	Z	p	q
1	AUTONOMIC NERVOUS SYSTEM IMBALANCE	0.1167	0.0043	0.1124	7.7076	0.0000	0.0000
2	ALANINE AMINOTRANSFERASE INCREASED	0.0455	0.0000	0.0455	4.9088	0.0000	0.0002
3	PYREXIA	0.0989	0.0297	0.0692	4.4888	0.0000	0.0009
4	HEPATIC ENZYME INCREASED	0.0455	0.0021	0.0414	4.4422	0.0000	0.0009
5	BODY TEMPERATURE INCREASED	0.0613	0.0170	0.0443	3.6283	0.0003	0.0216
6	ASPARTATE AMINOTRANSFERASE INCREASED	0.0198	0.0000	0.0198	3.1940	0.0014	NS
7	BLOOD POTASSIUM INCREASED	0.0138	0.0000	0.0138	2.6643	0.0077	NS
8	BLOOD PRESSURE INCREASED	0.0138	0.0000	0.0138	2.6643	0.0077	NS
9	HYPOTENSION	0.0237	0.0043	0.0195	2.6325	0.0085	NS
10	PLATELET COUNT INCREASED	0.0119	0.0000	0.0119	2.4642	0.0137	NS
11	DEBRIDEMENT	0.0138	0.0021	0.0137	2.3059	0.0211	NS
12	HYPERKALAEMIA	0.0138	0.0021	0.0137	2.3059	0.0211	NS
13	VOMITING	0.0119	0.0340	-0.0221	2.2934	0.0218	NS
14	ANAEMIA	0.0237	0.0064	0.0174	2.2558	0.0241	NS
15	HAEMOGLOBIN DECREASED	0.0178	0.0043	0.0136	2.0539	0.0400	NS
16	BLOOD ALKALINE PHOSPHATASE INCREASED	0.0000	0.0085	-0.0085	2.0086	0.0446	NS
17	OPEN WOUND	0.0040	0.0170	-0.0130	1.9813	0.0476	NS

#	Study Related Grouping (SRG)	ID (trt)	ID (plb)	RD	Z	p	q
1	LIVER FUNCTION TEST ABNORMAL	0.1305	0.0085	0.1220	7.8397	0.0000	0.0000
2	AUTONOMIC NERVOUS SYSTEM IMBALANCE	0.1167	0.0043	0.1124	7.7076	0.0000	0.0000
3	FEVER	0.1642	0.0510	0.1132	5.8538	0.0000	0.0000
4	HIGH POTASSIUM	0.0297	0.0021	0.0275	3.5143	0.0004	0.0124
5	HIGH BLOOD PRESSURE	0.0237	0.0043	0.0195	2.6325	0.0085	NS
6	TREATMENT PROBLEM	0.0237	0.0043	0.0195	2.6325	0.0085	NS
7	PLATELET COUNT INCREASED	0.0119	0.0000	0.0119	2.4642	0.0137	NS
8	LOW BLOOD PRESSURE	0.0257	0.0064	0.0193	2.4374	0.0148	NS
9	HIGH BLOOD GLUCOSE	0.0455	0.0212	0.0243	2.1280	0.0333	NS
10	HEARING IMPAIRED	0.0079	0.0000	0.0079	2.0080	0.0447	NS
11	DISORIENTATION	0.0040	0.0170	-0.0130	1.9813	0.0476	NS

3. Lumping based on correlations (upper: PT; Lower: SRG)

#	Preferred Terms (PT)	ID (trt)	ID (plb)	RD	Z	p	q
1	AUTONOMIC NERVOUS SYSTEM IMBALANCE	0.1167	0.0042	0.1124	7.7076	0.0000	0.0000
2	PYREXIA	0.1009	0.0064	0.0945	6.8057	0.0000	0.0000
3	ALANINE AMINOTRANSFERASE INCREASED	0.0455	0.0000	0.0455	4.9088	0.0000	0.0001
4	BODY TEMPERATURE INCREASED	0.0514	0.0021	0.0493	4.9063	0.0000	0.0001
5	HEPATIC ENZYME INCREASED	0.0435	0.0021	0.0414	4.4422	0.0000	0.0008
6	DECUBITUS ULCER	0.0336	0.0000	0.0336	4.1942	0.0000	0.0021
7	ANAEMIA	0.0218	0.0000	0.0218	3.3333	0.0008	0.0461
8	ASPARTATE AMINOTRANSFERASE INCREASED	0.0198	0.0000	0.0198	3.1940	0.0014	NS
9	RASH	0.0198	0.0000	0.0198	3.1940	0.0014	NS
10	HYPOTENSION	0.0237	0.0021	0.0216	3.0464	0.0023	NS
11	CELLULITIS	0.0178	0.0000	0.0178	3.0271	0.0025	NS
12	WOUND COMPLICATION	0.0178	0.0000	0.0178	3.0271	0.0025	NS
13	HAEMOGLOBIN DECREASED	0.0178	0.0000	0.0178	3.0271	0.0025	NS
14	DEBRIDEMENT	0.0138	0.0000	0.0138	2.8511	0.0044	NS
15	BLOOD POTASSIUM INCREASED	0.0138	0.0000	0.0138	2.6643	0.0077	NS
16	BLOOD PRESSURE INCREASED	0.0138	0.0000	0.0138	2.6643	0.0077	NS
17	PLATELET COUNT INCREASED	0.0119	0.0000	0.0119	2.4642	0.0137	NS
18	HYPERGLYCAEMIA	0.0119	0.0000	0.0119	2.4642	0.0137	NS
19	PLEURAL EFFUSION	0.0119	0.0000	0.0119	2.4642	0.0137	NS
20	HYPERKALAEMIA	0.0158	0.0021	0.0137	2.3059	0.0211	NS
21	OSTEOMYELITIS	0.0099	0.0000	0.0099	2.2472	0.0246	NS
22	RBC SEDIMENTATION RATE INCREASED	0.0099	0.0000	0.0099	2.2472	0.0246	NS
23	WBC COUNT INCREASED	0.0099	0.0000	0.0099	2.2472	0.0246	NS
24	HYPONATRAEMIA	0.0099	0.0000	0.0099	2.2472	0.0246	NS
25	MUSCULOSKELETAL PAIN	0.0099	0.0000	0.0099	2.2472	0.0246	NS
26	HYPERTENSION	0.0099	0.0000	0.0099	2.2472	0.0246	NS
27	GGTINCREASED	0.0178	0.0042	0.0136	2.0539	0.0400	NS
28	STAPHYLOCOCCAL INFECTION	0.0079	0.0000	0.0079	2.0080	0.0447	NS
29	WOUND	0.0079	0.0000	0.0079	2.0080	0.0447	NS
30	BLOOD UREA INCREASED	0.0079	0.0000	0.0079	2.0080	0.0447	NS
31	ANXIETY	0.0079	0.0000	0.0079	2.0080	0.0447	NS
32	DYSPNOEA	0.0079	0.0000	0.0079	2.0080	0.0447	NS

#	Study Related Grouping (SRG)	ID (trt)	ID (plb)	RD	Z	p	q
1	LIVER FUNCTION TEST ABNORMAL	0.1305	0.0085	0.1220	7.8397	0.0000	0.0000
2	AUTONOMIC NERVOUS SYSTEM IMBALANCE	0.1167	0.0043	0.1124	7.7076	0.0000	0.0000
3	FEVER	0.1602	0.0467	0.1135	5.9770	0.0000	0.0000
4	HIGH POTASSIUM	0.0297	0.0021	0.0275	3.5143	0.0004	0.0174
5	SKIN COMPLICATIONS	0.0831	0.0318	0.0512	3.4847	0.0005	0.0174
6	ANAEMIA	0.0435	0.0106	0.0329	3.2165	0.0013	0.0342
7	TREATMENT PROBLEM	0.0237	0.0021	0.0216	3.0464	0.0023	NS
8	HAEMORRHAGE	0.0138	0.0000	0.0138	2.8511	0.0044	NS
9	LOW BLOOD PRESSURE	0.0257	0.0043	0.0215	2.8061	0.0050	NS
10	WOUND COMPLICATION	0.0336	0.0085	0.0251	2.7730	0.0056	NS
11	VASCULAR DISORDERS	0.0138	0.0000	0.0138	2.6643	0.0077	NS
12	HIGH BLOOD PRESSURE	0.0237	0.0043	0.0195	2.6325	0.0085	NS
13	INFECTION	0.0316	0.0085	0.0232	2.6136	0.0090	NS
14	PAIN	0.0415	0.0149	0.0267	2.5455	0.0109	NS
15	MENTAL DISORDER	0.0178	0.0021	0.0157	2.5080	0.0121	NS
16	INFLAMMATIONS	0.0119	0.0000	0.0119	2.4642	0.0137	NS
17	NEUROLOGIC DISFUNCTION	0.0119	0.0000	0.0119	2.4642	0.0137	NS
18	PLATELET COUNT INCREASED	0.0119	0.0000	0.0119	2.4642	0.0137	NS
19	RESPIRATORY COMPLICATIONS	0.0158	0.0021	0.0137	2.3059	0.0211	NS
20	SEPSIS	0.0237	0.0064	0.0174	2.2558	0.0241	NS
21	HIGH BLOOD GLUCOSE	0.0455	0.0212	0.0243	2.1280	0.0333	NS
22	URINARY DISORDERS	0.1562	0.1125	0.0437	2.0116	0.0443	NS
23	GASTROINTESTINAL OBSTRUCTION	0.0079	0.0000	0.0079	2.0080	0.0447	NS
24	HEARING IMPAIRED	0.0079	0.0000	0.0079	2.0080	0.0447	NS

CONCLUSIONS

Summary of the results

Grouping Method	Analysis Category	m	# p < 0.05 (%)	# q<0.05 (%)
No Grouping	Individual Events (IE)	864	11 (1.27)	1 (0.12)
PT Level Grouping	PT	372	17 (4.57)	5 (1.34)
	PT(Correlation)	462	48 (10.39)	7 (1.52)
Study Related Grouping	SRG	113	11 (9.73)	4 (3.54)
	SRG(Correlation)	158	32 (20.25)	6 (3.80)

Model results

Preferred Terms (PT)	POI		ZIP		NBIN	
	p	q	p	q	p	q
AUTONOMIC NERVOUS SYSTEM IMBALANCE	0.0000	0.0001	0.0000	0.0000	NS	NS
PYREXIA	0.0002	0.0156	0.0002	0.0153	0.0136	NS
ALANINE AMINOTRANSFERASE INCREASED	0.0000	0.0000	0.0000	0.0004	0.0000	0.0009
BODY TEMPERATURE INCREASED	0.0002	0.0156	0.0310	NS	NS	NS
HEPATIC ENZYME INCREASED	0.0000	0.0000	0.0000	0.0014	0.0000	0.0020
DECUBITUS ULCER	NS	NS	NS	NS	NS	NS
ANAEMIA	0.0002	0.0156	0.0332	NS	NS	NS

Study Related Grouping (SRG)	POI		ZIP		NBIN	
	p	q	p	q	p	q
LIVER FUNCTION TEST ABNORMAL	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
AUTONOMIC NERVOUS SYSTEM IMBALANCE	0.0000	0.0000	0.0000	0.0000	NS	NS
FEVER	0.0000	0.0018	0.0000	0.0007	0.0098	NS
HIGH POTASSIUM	0.0000	0.0000	0.0024	NS	0.0035	NS
SKIN COMPLICATIONS	NS	NS	NS	NS	NS	NS
ANAEMIA	0.0224	NS	NS	NS	NS	NS

- A SAS program is developed to analyze high dimensional sparse counts data using SAS macros and procedures
- A lumping method is proposed to combine data correlation and biological pathway information to reduce data dimension
- Statistical inference is adjusted for multiplicity by controlling FDR
- Regression models with count data (POI, NBIN, ZIP) are used to cross check the inference results based on RD/RR
- The proposed method is to increase the inferential power without losing clinical results interpretation

ACKNOWLEDGEMENTS

Author gratefully thanks Ms. Anne Horney (VA Cooperative Studies Program, Perry Point) for her technical support and Drs. Joseph Collins and Zhibao Mi (VA Cooperative Studies Program, Perry Point) for their statistical advice

REFERENCES

- David I Gregorio *et al*. Lumping or splitting: seeking the preferred areal unit for health geography studies *International Journal of Health Geographics* 2005, 4(6), 1-10
- Marjorie Solomon *et al.* From lumping to splitting and back again: Atypical social and language development in individuals with clinical-high-risk for psychosis, first episode schizophrenia, and autism spectrum disorders *Schizophrenia Research* 2011, 131, 146-151
- Eviatar Zerubavel *et al*. Lumping and Splitting: Notes on social classification *Sociological Forum*, 1996, 11(3), 421-433
- Radu Tunaru Hierarchical Bayesian models for multiple count data *Austrian Journal of Statistics* 2002, 31(2), 221-229
- Jeppe Bennekou Schroll et al. Challenges in coding adverse events in clinical trials: a systematic review *PloS One*, 2012, 7(7) e41174

SAS® GLOBAL FORUM 2016

IMAGINE. CREATE. INNOVATE.

LAS VEGAS | APRIL 18-21

#SASGF