

S14: Hypothesis-Free Search for Connections between Birth Month and Disease Prevalence in Large, Geographically Varied Cohorts

John Borsi - Data Scientist, IBM Watson Health

Disclosure

I am employed as a Data Scientist for IBM Watson Health

Learning Objectives

After participating in this activity, the learner should be better able to:

- Understand the connection between birth month and lifetime health
- Apply statistical techniques to find connections in large health care data sets

Of Birthdays And Big Data

What time of the year you're born has measurable impacts on *lifelong* health

Neurological

“Neonatal vitamin D status and risk of **schizophrenia**”

2010, Archives of General Psychiatry.

“Relative Immaturity and **ADHD**”

2014, Journal of Child Psychology and Psychiatry

“Timing of birth and risk of **multiple sclerosis**”

2005, BMJ

Reproductive

“Month of birth and **offspring count** of women”

2008, Human Reproduction

“The impact of maternal birth month on **reproductive performance**”

2010, Journal of Biosocial Science

Endocrine

“Association of **type 1 diabetes** with month of birth among US youth”

2009, Diabetes Care

Immune disorders

“Month of birth, vitamin D and risk of **immune mediated disease**”

2012, BMC Medicine

Overall Longevity

“**Lifespan** depends on month of birth.”

2001, Proceedings of the National Academy of Sciences

And more...

Boland, *et al*'s search of PubMed revealed 92 articles pertaining to
16 different conditions

Big observational data sets are an obvious
way to search for these connections
systematically

Case Study 1

On the Effects of Pre-Natal Super Bowl Exposure

"It's Just a Game: The Super Bowl and Low Birth Weight"

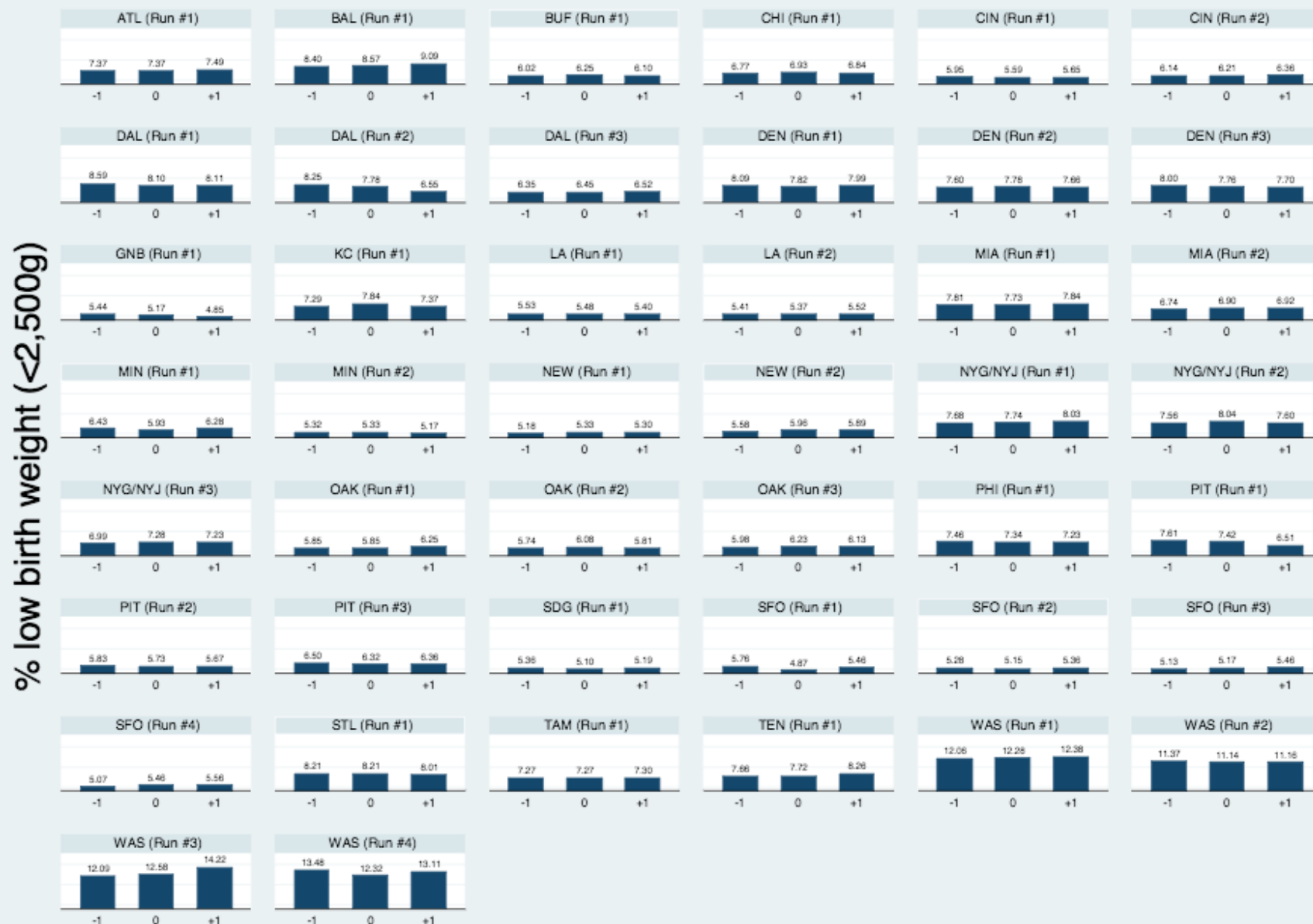
2016, Journal of Human Resources

29 Million Births over 40 years

National Vital Statistics System

Prenatal exposure to a super bowl increases
change of low birth weight ($p < .01$)

Figure 1: Percent low birth weight (<2,500g) babies in the years of and surrounding Super Bowl runs



Weaknesses of this approach

Overstated sample size

Hidden hypothesis-generation step

No statistical correction

"Our results suggest that informational campaigns aimed at encouraging pregnant women to avoid emotionally charged but otherwise ordinary events could lead to healthier babies"

Case Study 2

Season-Wide Association Study (*SeaWAS*)

Birth month affects lifetime disease risk: a phenome-wide method

Mary Regina Boland^{1,2}, Zachary Shahn³, David Madigan^{2,3}, George Hripcsak^{1,2},
Nicholas P Tattonetti^{1,2,4,5,*}

RECEIVED 7 January 2015

REVISED 23 March 2015

ACCEPTED 18 April 2015

PUBLISHED ONLINE FIRST 3 June 2015



















OXFORD
UNIVERSITY PRESS

1.7 M Patients over 28 years

Columbia University Medical Center

Lifetime risk of disease tied to month of birth

Table 2: Birth Month-Disease Associations Discovered Using SeaWAS (n = 16)

EHR Condition in SeaWAS	N	Passed Internal Validation?	Adjusted P^1	Seasonal Pattern	Birth Month Risk	
					High	Low
Cardiovascular (n = 9)						
Atrial fibrillation	48 961	Yes	<0.001		March	October
Essential hypertension	269 913	Yes	<0.001		January	October
Congestive cardiac failure	61 448	Yes	<0.001		March	October
Angina	20 741	Yes	<0.001		April	September
Cardiac complications of care	13 653	Yes	0.027		April	September
Cardiomyopathy	17 873	Yes	0.009		January	September
Pre-infarction syndrome	25 028	No	0.036		June	October
Chronic myocardial ischemia	10 010	No	0.022		April	November
Mitral valve disorder	22 966	No	0.024		March	November
Other (n = 7)						
Acute upper respiratory infection	112 487	Yes	<0.001		October	May
Bruising	8904	Yes	0.015		December	April
Nonvenomous insect bite	7435	Yes	0.001		October	February
Venereal disease screening	69 764	Yes	0.003		October	June
Primary malignant neoplasm of prostate	20 353	Yes	0.002		March	October
Malignant neoplasm of overlapping lesion of bronchus and lung	2714	Yes	0.014		February	November
Vomiting	30 495	No	0.029		September	January

Strengths of Boland, *et al*'s approach

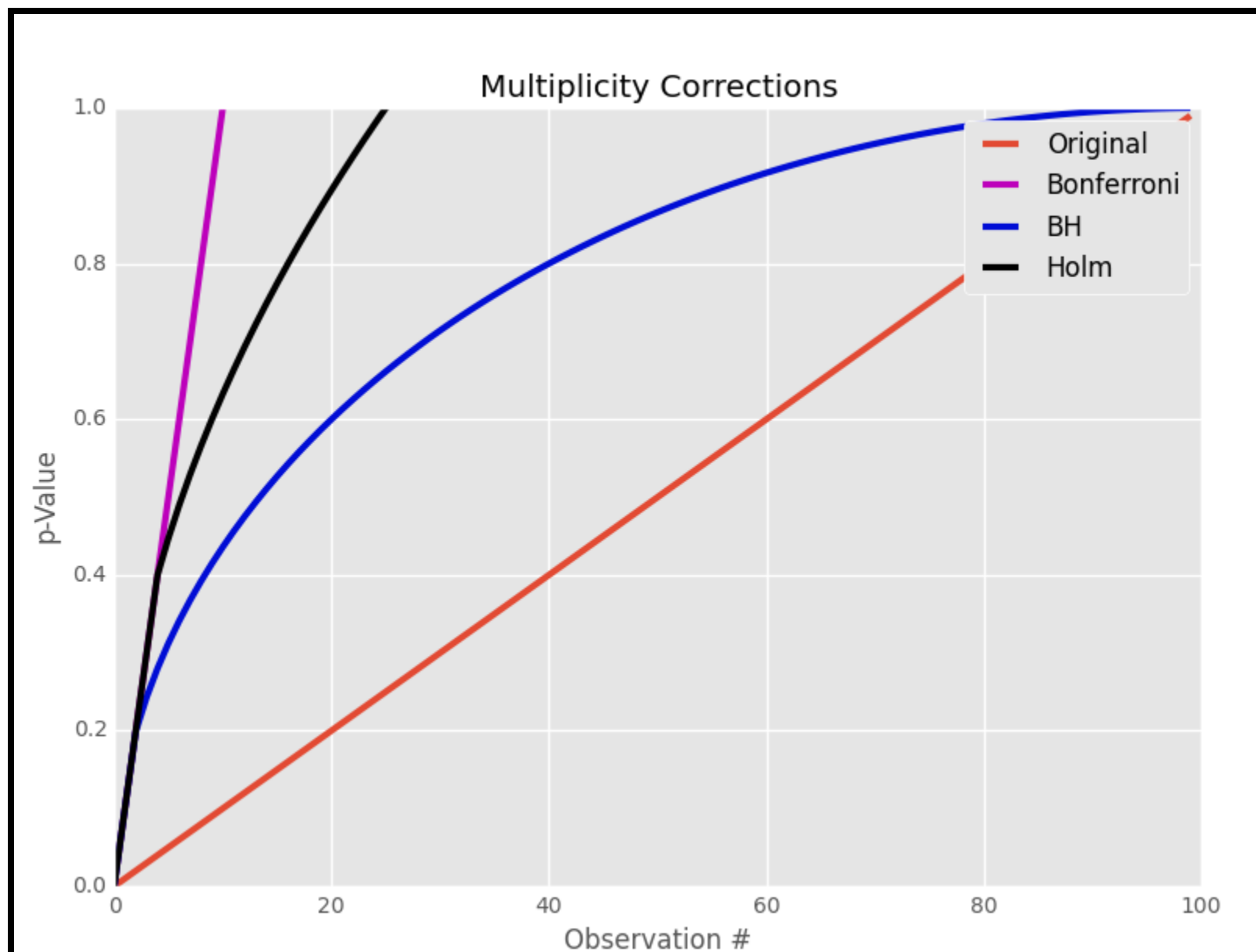
Accurate reporting of sample size

Upfront about hypothesis generation

Statistical correction

Limitations of Study

Sidebar: Dealing with Multiple Comparisons



Case Study 3

Multi-cohort Follow-Up to *SeaWAS*

"Hypothesis-Free Search for Connections between Birth Month and Disease Prevalence in Large, Geographically Varied Cohorts"

AMIA 2016

11.8 M Patients from multiple sources

IBM Explorys Life Sciences Data Set

- 3 distinct geographical regions
- Different types of data sources

Findings

- Geography and Data Source matter
- Holm multiplicity correction produces more reproducible results

Table 3. Comparison of results for Holm and Benjamini-Hochberg (BH) multiplicity corrections. SeaWAS results are given in columns labeled “NY”; replication results given in “C1,” “C2,” and “C3” columns.

Total Number of Associations Identified				
	NY	C1	C2	C3
Holm	-	39	16	5
BH	55	81	46	9
Number of Literature Supported Associations Identified [# (%Recall)]				
	NY	C1	C2	C3
Holm	-	9 (56)	7 (44)	1 (6)
BH	7 (44)	10 (63)	11 (69)	3 (19)

Number of Novel Associations Discovered				
	NY	C1	C2	C3
Holm	-	30	9	4
BH	16	71	35	6
Number of Novel Associations Validated in Other Cohort [# (%Precision)]				
	NY	C1	C2	C3
Holm	-	6 (20)	6 (67)	3 (75)
BH	4 (25)	8 (11)	14 (40)	4 (67)

"It is clear that this approach can be a powerful hypothesis generator and tool for investigating the role of seasonally dependent early developmental mechanisms in general health, but obtaining generalizable results requires evaluating different sets of data and accounting for potential biases."

Takeaways

Lifetime impact of pre-/peri- natal environment

Data source specific bias

Multiplicity corrections

Questions?

John Borsi - jborsi@us.ibm.com - @jpborsi