TOP TEN LIST

TEN COMMANDMENTS OF SECONDARY DATA ANALYSIS

With apologies to David Letterman, and thanks for editorial assistance to Elizabeth Kirby and Donna Petersen, and for their insights to the following colleagues:

Sheree Boulet, Emory University
Kimberlea Hauser, Pennsylvania Department of Health
Michael R. Kramer, Emory University
Lisa Marengo, Texas Department of Health Services
Bill Sappenfield, University of South Florida
J.P. Tanner, University of South Florida
Jason Salemi, Baylor College of Medicine

© R.S. Kirby, May 2019

Number 10

Thou shalt always obtain and use the data dictionary for the secondary dataset(s) thou intendeth to use.

Consider this documentation as thou wouldst a Rosetta stone.

Number 9

Thou shalt not ignore missing data*. This includes records that became 'missing' because they could not be linked with other data.

In fact, thou shalt strive to understand the factors that led to the state of missingness, and how those factors may influence thine results.

Paraphrasing Matthew 10:30, "Even the missing data is numbered."

Number 8

Always, thou shalt remember to fix your eyes on both the seen and unseen – just because something isn't reported in thine secondary data, doesn't mean it isn't so. And just because it is, does not mean it is so.

Using only hospital discharge data for women delivering during that hospital stay, many women living with disabilities or who are obese will not be so identified.

Corollary: You may not know *what* you don't know, but you should know of the *possibility* of not knowing it.

Number 7

Know thy data!

- Know how thy data was collected.
- Know its limitations, and state them, often.
- Know the quality of thine data and data elements.
- Be mindful of data conversion issues.
- And always, always conduct exploratory data analysis.

Number 6

Thou shalt not assumeth that bigger is necessarily better.

Large datasets don't necessarily provide valid results, even when all comparisons may be statistically significant.

Datasets with a multitude of variables can lead to multiple comparison bias and statistical significance due to chance.

Number 5

Know thine computational and statistical resources*.

Available computing power and statistical software may not suffice to process thine secondary data. Thusly, thou shalt learn ways to maximize efficiency in thine data management and analysis.

*This commandment can be paraphrased as:

Know thine limits.

Number 4

Know thy purpose!

And, stay true to it, even as thou produceth and discusseth thine analyses.

Number 3

Even as thou knoweth thy purpose, always remember:

The secondary data thou useth almost certainly were not collected with thy research project in mind.

Number 2

Though shalt not fish nor dredge thine data for research questions or hypotheses.

Secondary data sources can indeed support hypothesis generation, but thou shalt not then use the same data for thine analyses.

Number 1*

Thou dost not own thine secondary data.

Thou shalt always abideth by any restrictions made by the owner, especially concerning whether linkages to other data sources are permissible, data security requirements, and provisions for destroying the data after use.

Bonus Commandment

Thou shalt realize 'secondary' data does not imply 'secondary' importance or 'secondary' methodological rigor.

Analyses of secondary data can yield highly impactful findings. As such, as much (if not more) attention must be paid to every step of the epidemiological process.