Example of doing a data project

One of the things I did in my career to date was Director of Education for NAON, the National Association of Orthopedic Nurses (USA). During my tenure there, I became very interested in hip fractures and first learned that falls in the elderly often resulted in hip fracture, which was then associated with a poor prognosis, if not leading to death.

After speaking with some of you about the data project, it seemed that a strategy for the project might be useful. I thought about dividing the number of dataset search engines by the number of people working on a project, such that each person would take a subset of the search engines, then look for a topic of interest and find associated datasets. It seemed reasonable that a team of three, for example, could cover about 5 dataset search engines each and then look for a topic of interest and choose 2-3 datasets, then pool them with the rest of the team. This might result in 6-9 potential datasets, from which the team could choose the one or two that seemed to hold the greatest potential for responding to clinically meaningful questions. By examining these datasets, it is often easy to see variables that look like they might be related, either correlated or causally. From there, it is a simple matter to ask a question of the data and then analyze the variables of interest to see if a correlation or other relationship existed, such as differences between values of a variable based upon other characteristics of the subjects providing the data.

Let me use an example to demonstrate.

I am interested in hip fractures and falls in the elderly, so I chose Google Dataset search engine and entered ‘hip fracture’ as a search string:



Google presented suggestions, as I typed. I liked the sound of the second suggestion highlighted in green) so clicked on it. This took me to:



I clicked on the Figshare button to get to the dataset source and was shown:



I noticed at the bottom of the data that I could download the original file, so I did and got an Excel file, and here is a part of it:



I looked at the columns for each of the variables and noticed some interesting potential relationships between the BI (Barthel Index – green highlight) and the outcome (yellow highlight). I knew I could do a Spearman correlation analysis on those two variables. So, I wondered about the question I could ask that could be answered by transforming this data and moving along the DIKW continuum. One typical and potentially useful question I could ask would be something like “Is an unfavourable outcome (prognosis) related to the Barthel index?” A Barthel index is a standardized reliable clinical assessment that gives a numeric estimate of the disability of a fracture patient. My thinking was that it would be useful to know which patients were at higher risk following a fracture, since then we could think of stratifying care based upon degree of risk.

It was a simple matter in Excel to show that there was a correlation between the Barthel index and outcome by calculating Spearman’s rho. As I looked at other variable values, it was clear I could have examined relationships between several other variables.