Data Snooping Part Four

How's that classification scheme working out for you?

Donald J. Wheeler

In Parts One and Two on this series we discovered some caveats of data snooping. In Part Three we discovered how listening to the voice of the process differs from the model-based approach and how it also provides a way to understand when our models do and do not work. Here we conclude the series with a case history of how big data often works in practice.

Daniel Boorstin summarized the essence of distilling knowledge out of a database when he wrote: "Information is random and miscellaneous, but knowledge is orderly and cumulative." As we seek to organize our miscellaneous data we have to be careful to make a distinction between signals and noise. The following is the story of one attempt to turn data into knowledge.

The client in this story is a large health-maintenance organization. Their database lists every transaction between their members and healthcare providers in the state. At the end of each year these transactions are automatically gathered into "episodes" that track the treatment of a single patient by a single provider from the initial visit to the conclusion of treatment. For each episode several different characteristics are recorded. Of course, the HMO wanted to compare costs per episode between providers.

To facilitate this comparison a consulting firm organized these episodes into equivalence classes using the standard statistical techniques of classification analysis. This classification scheme grouped together episodes that were alike in each of several ways. Episodes placed in the same group shared the same diagnosis; had patients that were in the same age range; had comparable complications during the treatment episode; had the same accompanying illnesses; and involved the same treatments.

The resulting Episode Treatment Groups (ETGs) were intended to be equivalence classes where all the records in a given ETG would be logically comparable. These episodes within a single ETG could then be organized to make comparisons between providers in terms of other variables such as patient outcomes and costs.

ETG 678

To illustrate how this worked we shall use ETG 678 and provider 708. Provider 708 had 35 patients in ETG 678 during the course of the year. The average cost per patient for Provider 708 for ETG 678 was $2,149. This value allowed the HMO to rank provider 708 relative to all the other providers for ETG 678. Simple, easy, and wrong.

Figure 1 lists the total cost for each of provider 708's patients in ETG 678. These values are arranged in columns in order of the starting date for each episode.

260	293	93	207
130	195	278	153
189	571	185	209
1,080	55,698	123	243
175	209	9,434	110
200	1,825	408	306
193	239	570	343
120	290	118	244
33	254	238	

Figure 1: Total Costs for ETG 678 Episodes for Provider 708

Even a cursory examination of the values in Figure 1 will show that they cannot be considered to be similar. These 35 values are clearly a mixture of apples and oranges. The average value may be $2,149, but 33 of these 35 values fall below this average! Using the average to summarize these 35 values is completely inappropriate. Eighty-six percent of the total cost for all 35 episodes comes from two episodes. A slight change in the cost of either of these episodes could easily change the ranking for provider 708. (Even simple descriptive statistics implicitly assume homogeneity for the data being summarized.)

A much better measure of location for these data would be the median cost of $238. This median value was used, along with the median moving range of $125, to create the *X* chart shown in Figure 2.

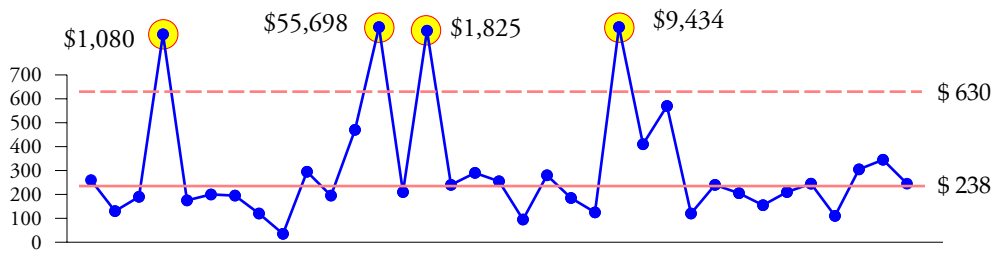

Figure 2: Total Costs for ETG 678 Episodes for Provider 708

Four of these episodes stand out as detectably different from the others. These four episodes do not belong in the same group with the remaining 31 episodes. Even though these 35 episodes are similar in so many ways, they are not homogeneous enough to be characterized by a descriptive statistic like the average cost. Clearly the classification scheme that created this episode treatment group was unsuccessful in creating a homogeneous grouping.

So, is this problem unique to provider 708? Not at all. When a chart like Figure 2 was created for those providers that had at least 20 patients in ETG 678, each chart contained between 1 and 8 exceptional episodes. Figure 3 lists these providers, their number of episodes, the number of episodes with excessive costs, and the proportions of excessive cost episodes.

Provider ID	**104**	**140**	**147**	**376**	**378**	**411**	**564**	**643**	**664**	**668**	**698**	**703**
No. Cases	30	40	37	34	20	30	40	21	21	24	30	29
No. Excessive	8	7	5	7	2	5	4	3	5	3	2	8
Proportion	0.27	0.18	0.14	0.21	0.10	0.17	0.10	0.14	0.24	0.13	0.07	0.28

Provider ID	**704**	**706**	**708**	**714**	**719**	**725**	**740**	**741**	**747**	**776**	**977**	**1021**
No. Cases	22	29	35	22	23	32	34	24	20	29	20	37
No. Excessive	4	4	4	4	2	5	3	1	1	3	1	4
Proportion	0.18	0.14	0.11	0.18	0.09	0.16	0.09	0.04	0.05	0.10	0.05	0.11

Figure 3: Proportions of Episodes with Excessive Costs for 24 Providers for ETG 678

A total of 95 out of these 685 episodes were found to be excessive (13.9 percent). The average moving range for these proportions is 7.22 percent. Figure 4 shows the 0.05 ANOX chart for the proportions of episodes with excessive costs by provider. (For more about the ANOX chart see Wheeler and Beagle, *Quality Digest*, Sept. 4, 2017.)

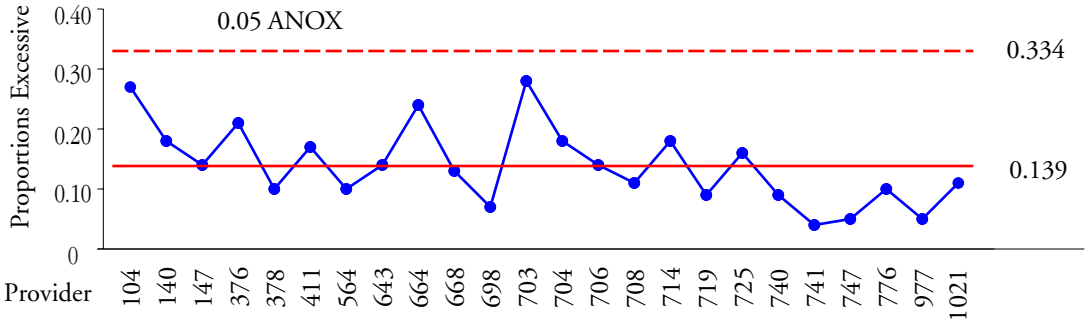

Figure 4: Proportions of Episodes with Excess Costs for ETG 678

There is no detectable difference between these 24 providers in terms of the proportion of episodes with excessive total costs. While the percentages of excessive costs range from 4 percent to 28 percent, these percentages are not sufficiently different to single out any provider as being different from the others. A similar analysis on the median costs per provider resulted in another chart like Figure 4 where *no one provider could be singled out from the rest*.

OTHER ETGs

Moreover, when this whole analysis was repeated for several other ETGs we found them to also be internally non-homogeneous. In other words no ETG was found to display the homogeneity necessary to allow meaningful comparisons to be made between *average* costs for providers within that ETG. Since these comparisons were the purpose behind the creation of the ETGs, this lack of homogeneity undermined the analysis used by the HMO.

So what can the HMO do? Let us assume that they will still want to make comparisons between providers within ETGs. As may be seen in Figure 2, they might use medians rather than averages to summarize location. (While this does not remedy the lack of homogeneity, it does make the analysis more robust to the influence of the extreme values.) Then, as seen in Figure 4,

they could filter out the noise between providers by using an ANOX chart to compare the providers within a particular ETG. With these two fixes they could avoid making serious errors when comparing providers within an ETG.

WHAT HAPPENED?

While the classification scheme successfully sorted the episodes according to the five variables listed above, there are other important variables that affect the cost of an episode that were not considered in the classification. Whether these important variables were missing from the database, or were overlooked in the analysis, we cannot tell. But when we inspect the results of our analysis (as suggested in Part Three of this series), we find a lack of homogeneity within each ETG which tells us that this classification scheme has failed to accomplish the client's stated objective of creating uniform groups to use in comparing providers.

THE CLASSIFICATION FALLACY

The classification scheme used by the consulting company to organize this database failed to create homogeneous groups because that is not what a classification model does. Classification models identify groups that are *different* from each other. They group together records that have similar values along each of several different axes (where an axis may consist of a single variable or a linear combination of several variables). In this way records that are in different groups will be different from each other on one or more axes. Moreover, records that are in the same group will have *similar* values on those axes used by the classification scheme. But records in the same group may still differ from each other on *other* axes.

The classification scheme above identified groups that *differed* in terms of diagnosis, age of patients, complications encountered, accompanying illnesses, and treatments used. But being *different* in these ways does not assure that the resulting groups are *homogeneous* in other ways.

Episodes in a single ETG will be *similar* in diagnosis, age of patient, complications encountered, accompanying illnesses, and treatments used. Yet these episodes may *differ* in many other and possibly important ways. In fact, in every ETG studied, they did differ. So while the classification scheme that created the ETGs may have organized a vast database into manageable groupings, and while these groupings may be more homogeneous than the database as a whole, this does not, in and of itself, justify comparisons *within* a grouping.

Thus, the "classification fallacy" is thinking that because you have reduced a database into a set of groups that differ from each other along a set of axes, each of these groups will then be internally homogeneous. This is rarely the case. In fact, not only may records in the same group differ from each other on other axes, but they may also, to a lesser extent, differ from each other on those axes used to define the groupings.

SUMMARY

Big data approaches seek to distill knowledge out of information by organizing our data or by finding models for the relationships between certain variables in a database. The data are input and following some computerized smoke and mirrors the results appear like magic. Yet regardless of the illusion of being the result of a rigorous analysis, these results still have to be

interpreted in light of the principles of data analysis.

The first principle of data analysis is that data have no meaning apart from their context. It is the context that will tell us where to expect homogeneity (within providers) and where we might expect to find differences of interest (between providers).

The second principle is that when we look for these differences we have to first filter out the noise. If we find a lack of homogeneity where we did not expect it (within providers) then we know that either we have not organized the data appropriately or else there are dominant assignable causes that we have overlooked. If we fail to find any differences where we expect to find them (between providers) then any apparent differences are merely noise and cannot be trusted to be real effects.

Thus homogeneity is not only the primary question of data analysis, but it can also be the key to properly interpreting the results of your data snooping efforts.