Paper 376-2008

Using Regression Diagnostics and Clustering Criteria to Uncover ID Errors

David J. Pasta, ICON Clinical Research, San Francisco, CA
Eric Elkin, ICON Clinical Research, San Francisco, CA
Rory Michaelis, ICON Clinical Research, San Francisco, CA

ABSTRACT

In a large observational study, key patient information was not always recorded accurately at every visit, including the patient identification number, the site, and the date of the visit. Because of the nature of the study, some patients were seen at more than one site the same year (possibly unknown to the other site) or might be seen for some years at one site, drop out, and reappear at the same or a different site some years later. In addition, there are patient ID numbers that appear to be in use by two different individuals simultaneously. With 13 years of data collection and tens of thousands of individuals, we needed an efficient way to detect the most obvious of the ID errors. By evaluating diagnostics resulting from running PROC REG and PROC CLUSTER on height and lung function data, we were able to identify situations where ID errors were likely. This paper steps through the process of developing the criteria we used; perhaps just as important, it shows some of the things we tried that turned out not to work well at all.

INTRODUCTION

Data had been collected on tens of thousands of patients in an observational study of cystic fibrosis (CF) over a period of 13 years. Cystic fibrosis is a genetic disease that affects lung and digestive functions and generally leads to a decline in health over time and premature death. There are a limited number of CF specialty clinics in the United States, and a large fraction of them participated in the study. This meant that the same patient might be seen at different sites sequentially (transferring care from one to another), but it was also possible for a patient to be seen at two different sites for multiple years. This could occur, for example, when a patient was receiving care at one center while away at college but at another center while home for the summer. Also, it is possible for a patient to drop out of the study when transferring to a center that did not participate and then reappear in the study some year later after transferring to another participating site.

The design of the study allowed for these possibilities. Specifically, there was no requirement that a patient be associated with a single site at any given point in time. This meant that an error in recording the patient identification number or the site number would be more difficult to detect because there was no definitive cross-reference between sites and patients. It also meant it was possible for two different patients to have their data recorded under the same patient ID number at the same time. Furthermore, if the site number was entered incorrectly a query might be sent to a site with no knowledge about that patient and therefore no ability to correct the error.

Although some sites entered data as patient visits took place, other sites "batched" their data entry and performed it quarterly or even annually after the end of the calendar year. This led to a tendency to record the year of the visit incorrectly. When entering data in February 2003 about a visit in January 2002, one would be sorely tempted to enter the current year (2003) for the visit instead of the correct year (2002). This type of year error could be introduced either by the site personnel filling out the form or by the data entry operator keying it into a database. An error in even a few percent of the visits could lead to hundreds or thousands of one-year date errors.

Because of the nature of the study, only limited resources were available to identify and correct data errors as data were collected. If the study were smaller, or if the data checking were undertaken while the data were being collected, it would be possible to examine all the suspicious data individually. We were faced with the task of analyzing the data without the ability to query the sites about possible errors. With so many patients with so much data, we could well afford to leave out some visits or even some entire patients if we had doubts about the validity of the data. But how could we identify the questionable data? It would have been prohibitively expensive to review every patient's data in detail "by hand." Was

there something we could do that was more automated? If the automated approach removed some potentially accurate data along with the highly questionable data, that would be acceptable. If it identified a small number of situations where it appeared there was a systematic sharing of an ID number between two different individuals, we could afford to investigate those in detail and perhaps correct the error. In short, we wanted to try to statistically clean the data and identify the "low-hanging fruit" of potential ID errors worthy of individual attention.

AN INITIAL APPROACH, USING HEIGHT DATA

For CF patients, height data is recorded very frequently. Because CF generally affects digestion it also affects height and weight, with CF patients much smaller in stature and weight than their peers. For this reason, height and weight are closely monitored and are generally checked at every office visit. In addition, lung function is routinely checked using pulmonary function tests. Because the normative tables for lung function rely on height to estimate the expected lung capacity, height is also recorded at each test. Unlike weight, which can fluctuate up or down, height generally rises steadily over time for children and young adults with decreases in height attributable primarily to measurement, transcription, or data entry error. Thus it appeared that evaluating height information over time would be a good way to find some of the most obvious ID errors. A patient who suddenly appeared to be much shorter or much taller could possibly be another patient altogether.

We had heights recorded in centimeters and the date of the associated visit or pulmonary function test. Initially, we thought that all we needed to do was look for big changes in height between adjacent measurements of height. We assumed that two people measuring the height of a child might differ by 1-2 centimeters, so we decided to look at cases where the height dropped by at least 3 centimeters. (An increase that large might represent a growth spurt, so we focused on decreases.)

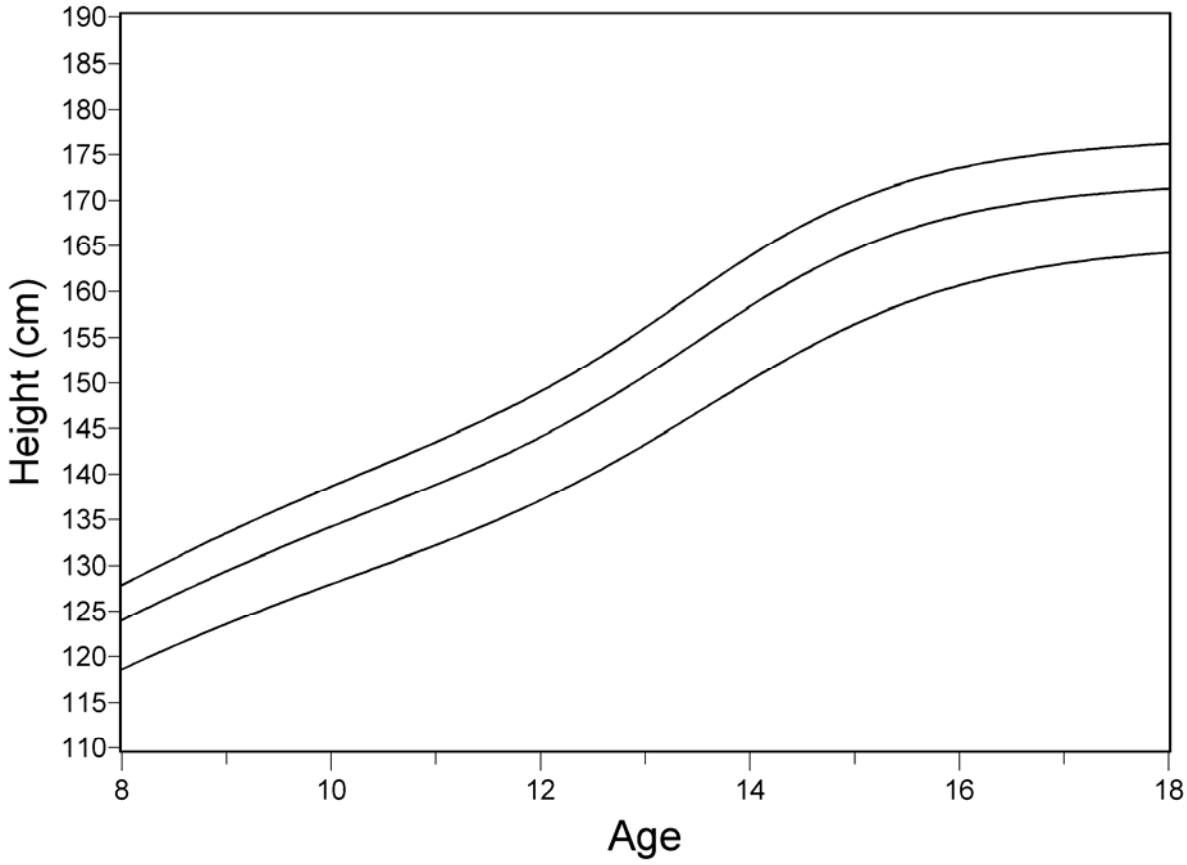
We were looking for low-hanging fruit, but what we found was an entire orchard. Thousands of patients had at least one 3 cm drop in height for adjacent measurements. So we looked for at least a 5 cm drop. Still too many. How about 10 cm? That's 4 inches – surely that's a big change in height. Yes, the number of potential problems was much less using that criterion, but still too many to look at individually. We needed to think more about what was going on.

We graphed the data for some of the patients and began to see some patterns. In many cases the big drop in height was preceded by an equally big increase. For example, a patient might have heights 135, 133, 136, 173, 138, 141. You and I look at that sequence and we immediately see that the 173 probably should be a 137 – a simple transposition of digits. When we look at differences from the previous measurement, it's a 37-cm increase in height followed by a 35-cm decrease. Another sequence might be 135, 133, 136, 138, 141, 140, 143, 147, 137, 152, 150, 151. The 137 is picked up by the algorithm – a 10 cm decrease followed by a 15 cm increase – but it turns out if you look in detail at the data the 137 would fit nicely into the growth curve if you moved it one year earlier. It's an error, all right, but it's an error in the date, not in the ID.

GETTING A LITTLE MORE SOPHISTICATED ABOUT HEIGHT

Maybe we should look at both increases and decreases in height – but pay attention to how rapidly those changes took place. That way we could detect those situations where there was one wildly high or low value in the middle of a reasonable series of heights. It helps to know that growth generally occurs in spurts, with measurable growth of up to 3/8" occurring literally overnight. It is possible for children to grow several centimeters in just a few weeks, so any criteria for change scores need to allow for substantial variation. We tried various criteria for picking out a big increase or decrease in height, taking into account the amount of time since the last measurement. We found that when there was a big gap in the data – several years, for example – we had a tendency to miss potential errors. For example, we allowed children to stay the same height for many years even though it would be very unusual for a 10-year-old to be the same height as they were at age 6.

So what might be reasonable amounts of growth over various periods of time? There are standard growth curves for the US population, giving percentiles of height by age separately by sex and race/ethnicity. The growth curves for the 50th, 25th, and 5th percentiles for children aged 8 to 18 as reported by the National Center for Health Statistics provide some guidance (see figure). Maybe we could use the growth curve information to track the percentile (or z-score) of patients over time?



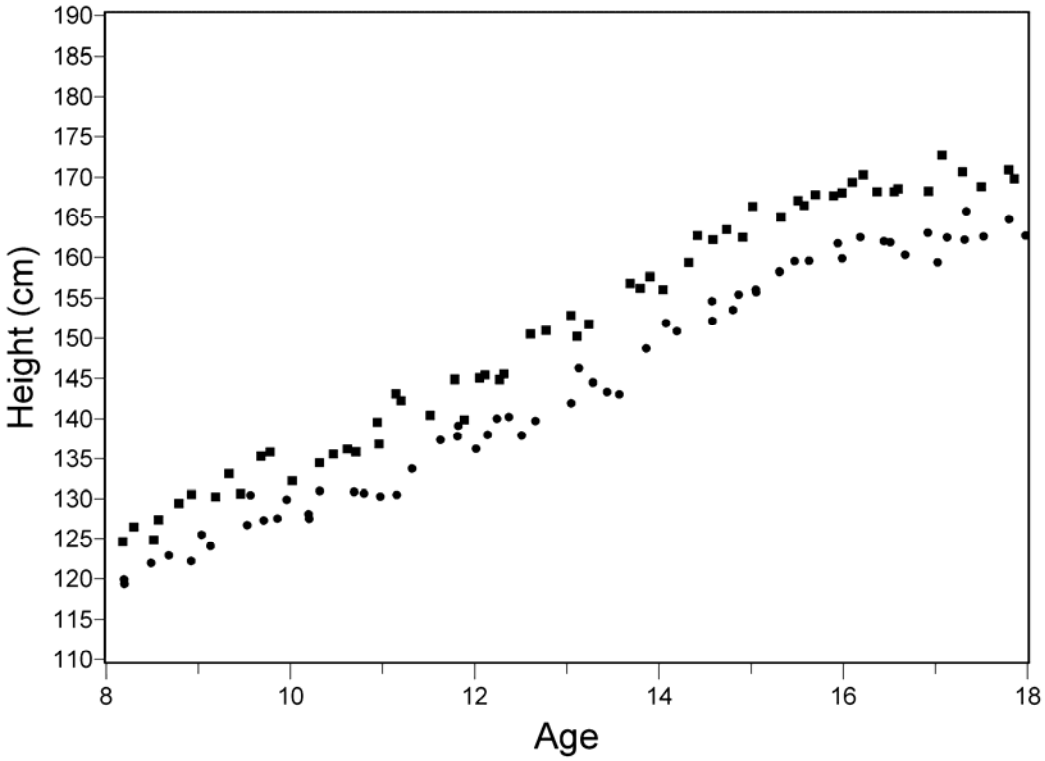
Because CF patients tend to be small – sometimes very small – a substantial fraction of them are below the 1st percentile in standard tables. For that reason, it makes more sense to work with z-scores than percentiles. Z-scores might be reasonably stable over time in the absence of errors.

The graphs of patients' z-scores for height were more stable than the actual heights, but they were no longer monotonically increasing. CF patients can and generally do lose ground relative to the general population, resulting in z-scores that trend downward over time. How much the z-score might reasonably drop over time depends in part on the expected growth in the general population. In contrast, it is possible to set reasonable (albeit broad) limits on the amount of growth in centimeters that is reasonable in a period of time, and of course no material drop in height is reasonable. For these reasons, we decided to go back to actual height rather than z-scores relative to the general population. With our new approach we were able to identify implausibly rapid growth over a short time as well as implausible decreases in height. But again often the errors appeared to be simple errors in the recording of heights or dates.

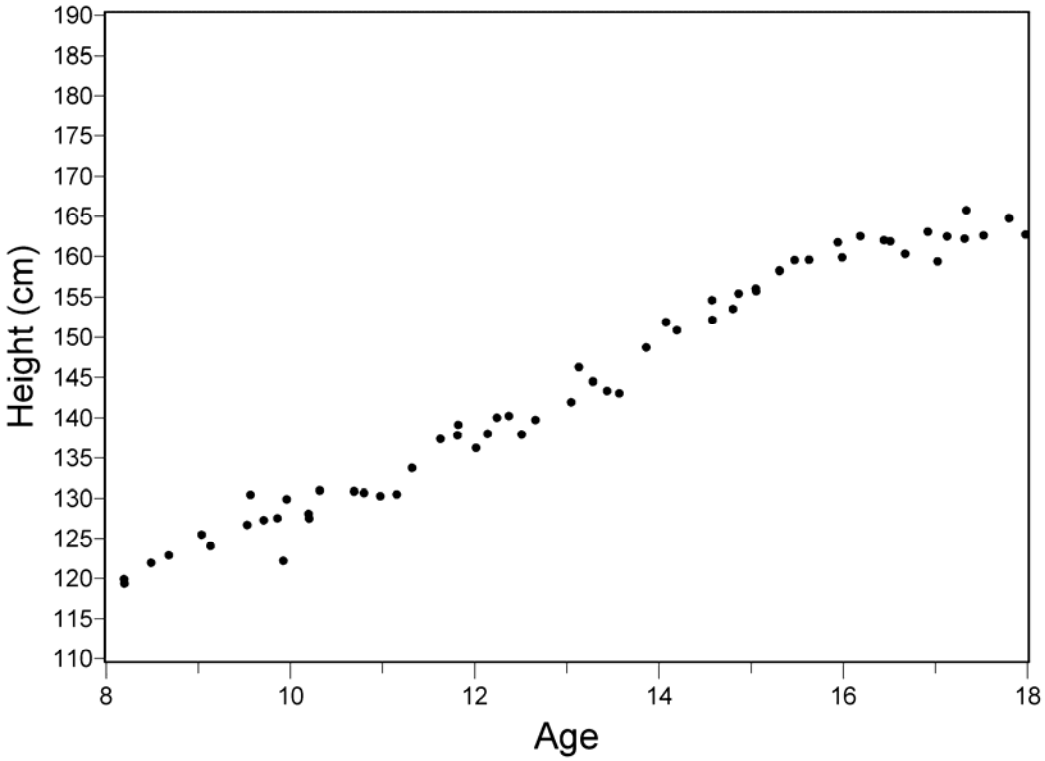
IF YOU KNOW WHAT YOU'RE LOOKING FOR, YOU KNOW WHEN YOU'VE FOUND IT

Finding anomalous heights was relatively easy, but it didn't accomplish our purpose. We needed to be able to accommodate the occasional errors in height or date in our statistical analysis. What we were trying to do was detect ID errors. How could we separate the two? What pattern of heights represented a potential ID error and what pattern represented a different type of error? To better understand that, we considered a number of different patterns that we had seen in the data and tried to characterize which ones represented ID errors and which ones did not.

Simulated data are presented for two hypothetical male patients, one with stature around the 25th percentile of height and one with stature around the 5th percentile of height are shown in the following figure:

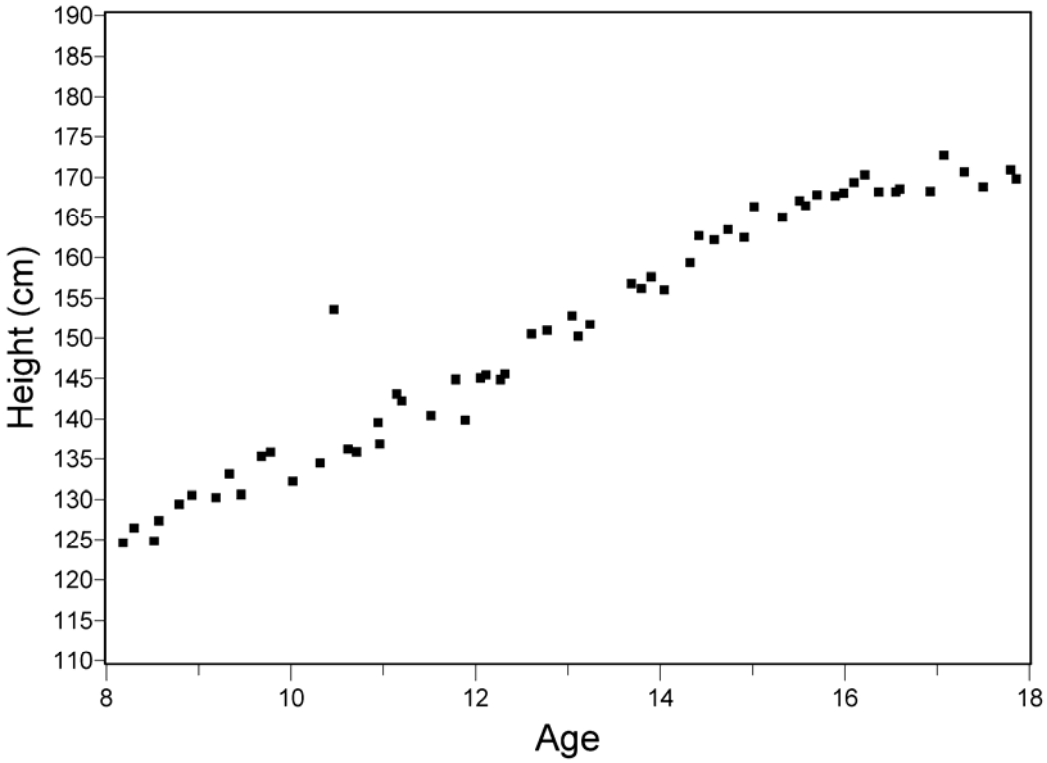


One error pattern that was observed is illustrated below:



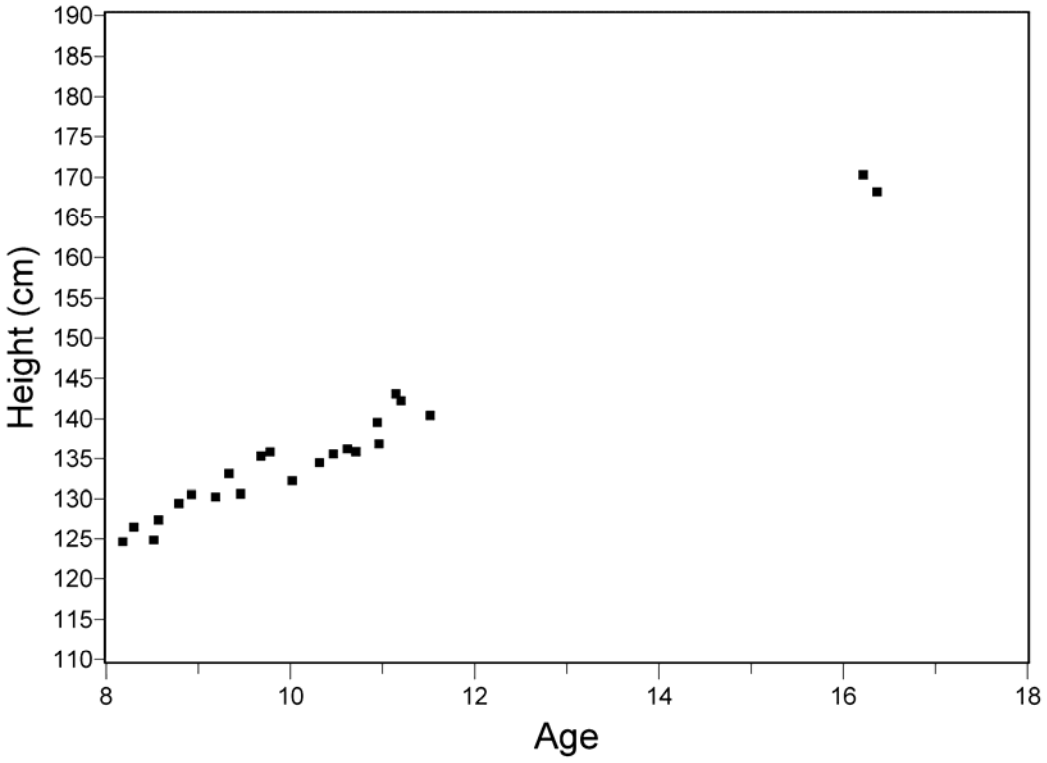
At about age 10, there seems to be a non-negligible decrease in height. However, if the data point were "back-dated" one year, it would fit into the overall pattern for this patient. This is an illustration of a date error.

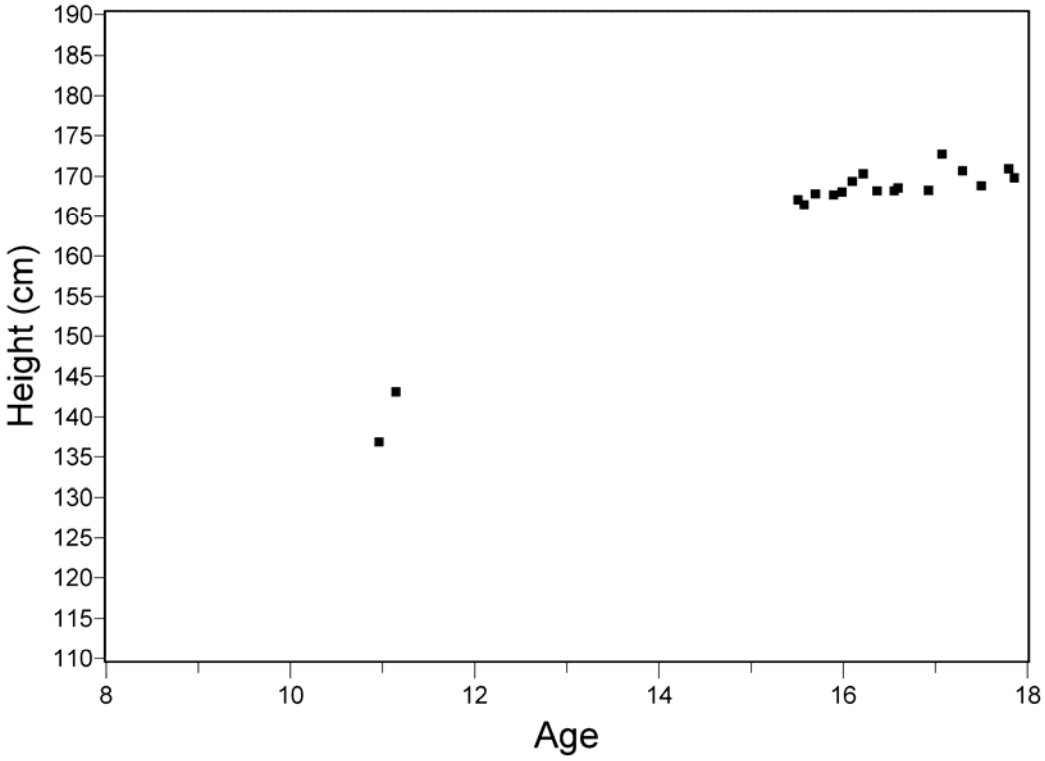
Another error pattern that was observed is illustrated below:

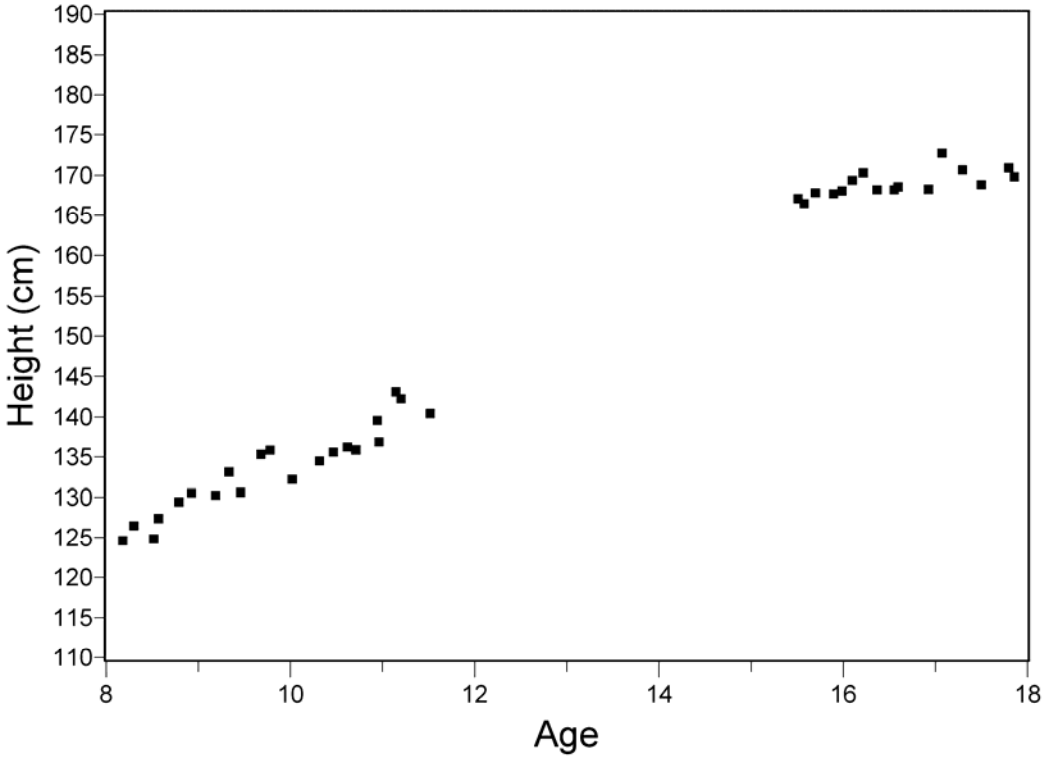


Here, a little after age 10 there is a large outlier. The height for this data point is 153.6 rather than 135.6. This is an example of a transposed digits error.

Another pattern which was not uncommon to see in the data is a large gap between visits (i.e., a gap up to several years) as shown in the next three graphs:

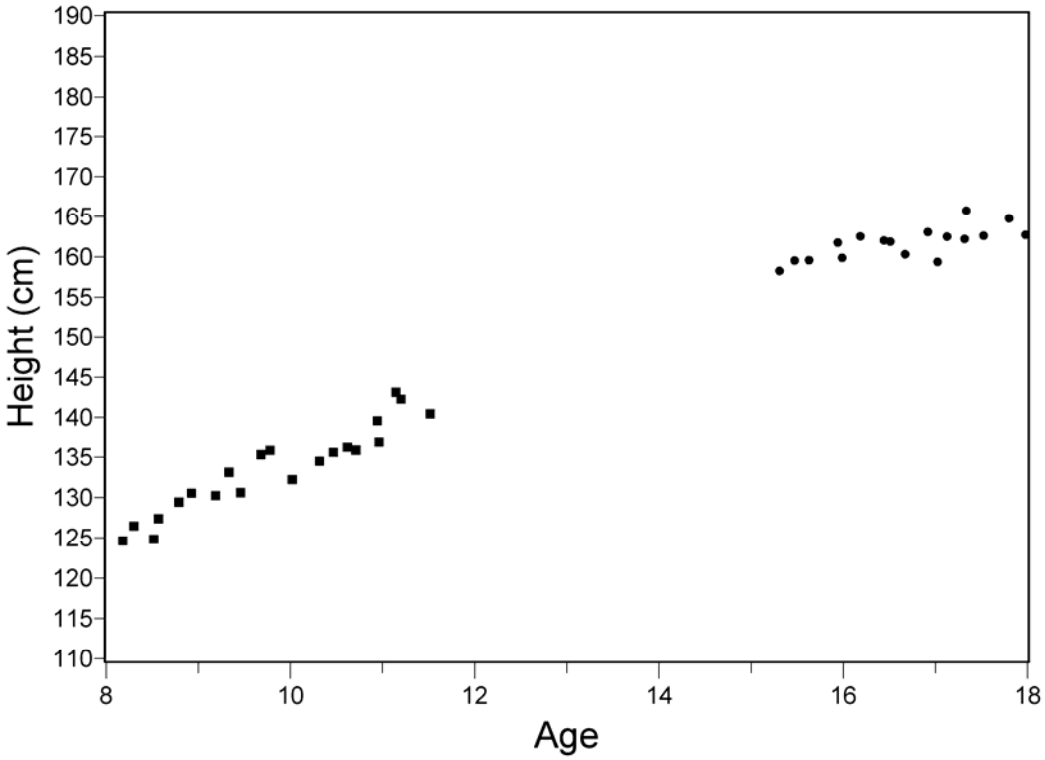


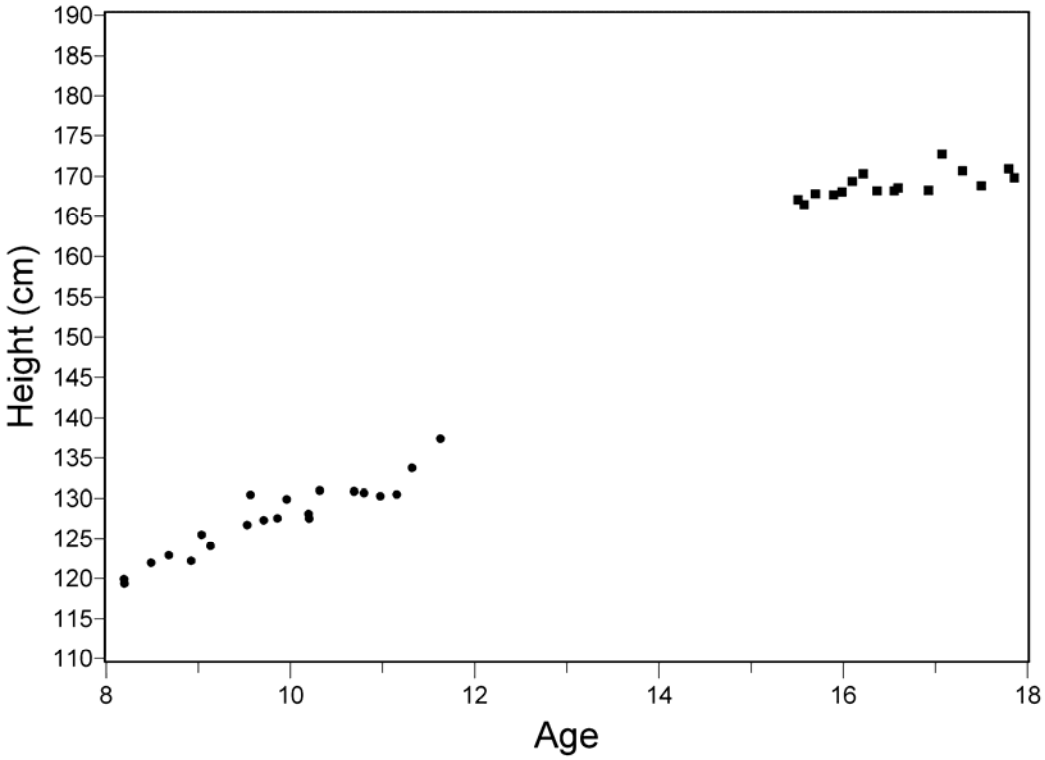




In the first of these graphs, it's possible a patient has returned after a long absence (and more data would have been forthcoming if study data collection had not ended). The second graph represents a more uncommon situation. In both cases, it is hard to tell if the two points which are far away from the main cluster of measurements are correct or belong to a different patient. In the third graph, they do not appear to be isolated ID errors because there are so many of them. However, they may or may not be the same patient. In this case, the data are all from the same hypothetical patient.

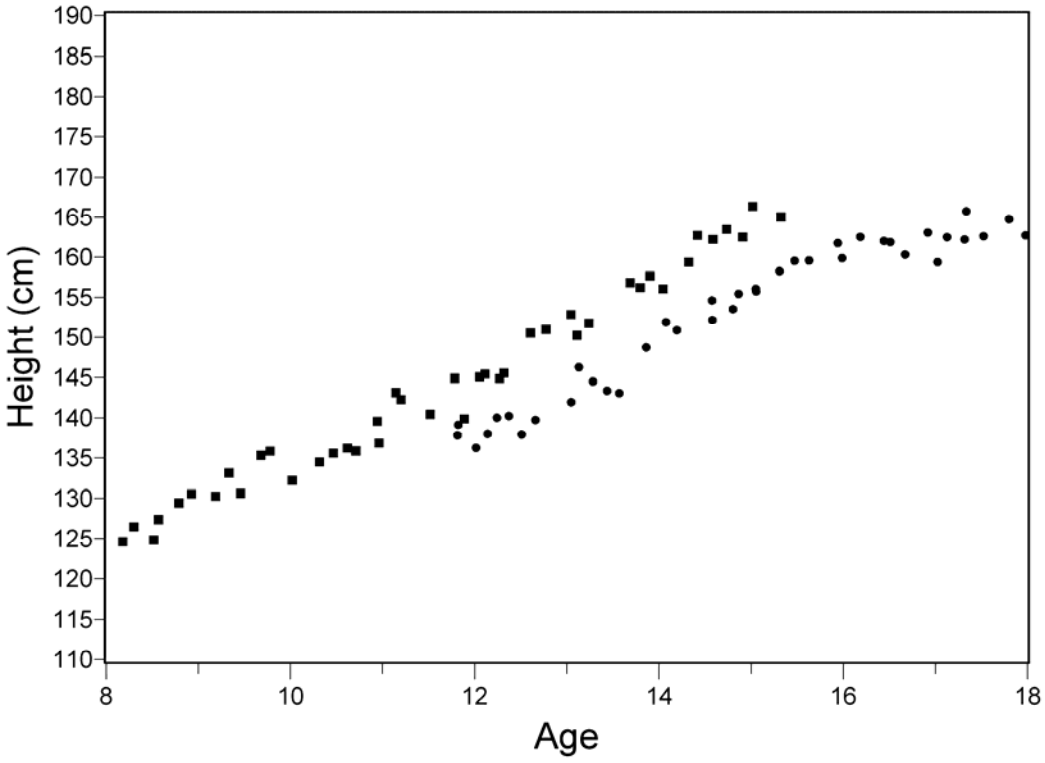
The next two examples show data combined from the two different hypothetical patients:



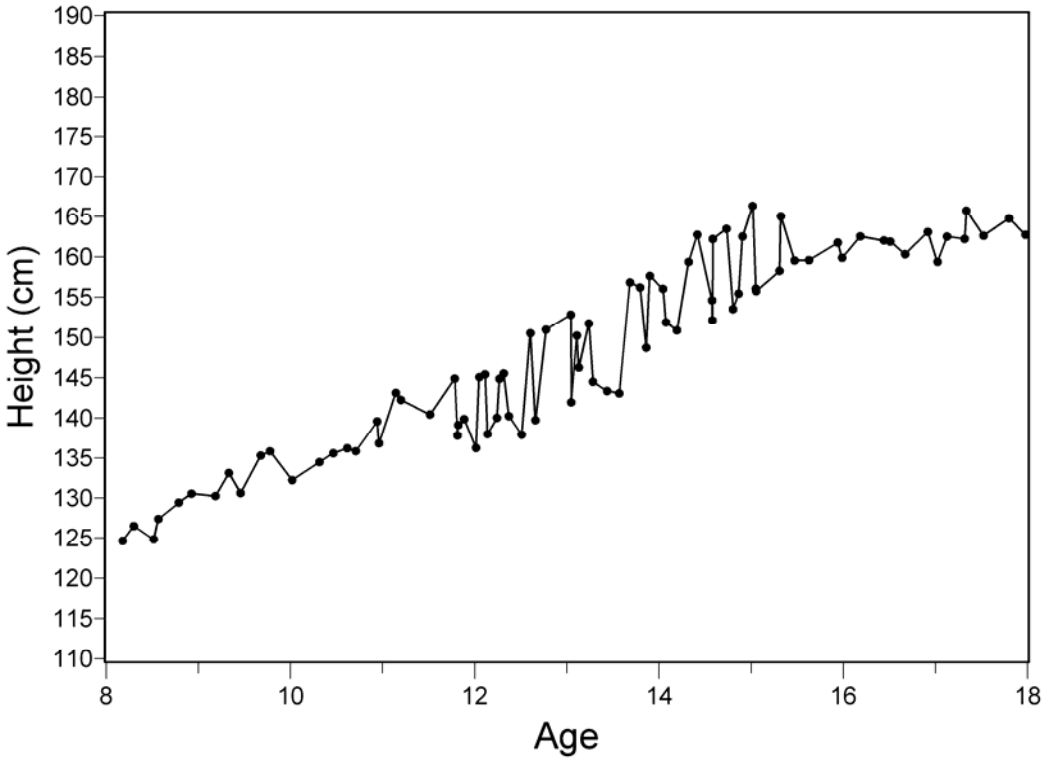


In these graphs, it is again challenging to decide if the data were from one patient or two. The first figure gives quite a reasonable pattern for a CF patient – to have dropped from the 25th to the 5th percentile. The second gives a less likely scenario – an improvement in percentile – but it is possible.

Finally, one more example will illustrate another pattern that is observed and which more clearly points to two different patients sharing the same patient ID number. If we plot the data as the two different patients we know they are, it would look like the following:



However, if we did not know in advance that it was two different patients, then the plot would look like the following "sawtooth" pattern, indicating the data were likely for two different patients:



Now that we (think we) know what we're looking for, how can we use statistics to help us find it?

LOOKING FOR BIMODALITY

Based on the pictures we looked at, we wanted to find situations where the heights appeared to represent two different distributions – not just a distribution with a single outlier in either the X or Y direction. We also couldn't assume the height would be a constant (even with z-scores), so we needed to look at residuals from a model. Although curvature is evident in the growth of children, "to a first approximation, everything is linear." (To a second approximation, everything is quadratic. To a zeroth approximation, everything is constant.) If we looked at residuals from a linear fit and there were clearly two different individuals combined, we should see two clusters of residuals (that is, bimodality). We could probably sharpen the contrast by fitting a more sophisticated model that included curvature, but we were looking for the low-hanging fruit.

We thought of examining bimodality by using the CLUSTER procedure in SAS® to evaluate the approximate number of clusters in the residuals. There are various criteria that can be used in connection with PROC CLUSTER to estimate the number of clusters in the data. Depending on the clustering technique used, different types of clusters are likely to be found. It also turns out there is a bimodality coefficient available from PROC CLUSTER.

Unfortunately, once again we found there was too much signal – lots of situations where there was evidence of multiple clusters in the data but that turned out not to be of particular interest. We tried different criteria for deciding on the number of clusters: the Cubic Clustering Criterion or CCC, the pseudo F, and the pseudo t^2. We tried different clustering techniques, although we emphasized Ward's method, which basically minimizes the within-cluster variance. This tends to produce clusters of similar size rather than picking out a few outliers – which is what we wanted. But it is very sensitive to outliers. Details on the clustering methods and the criteria for determining the number of clusters can be found in the chapter of the SAS/STAT manual on the CLUSTER procedure.

WE KEEP FINDING THOSE OUTLIERS

As we reviewed the graphs of height by age for the patients identified by the PROC CLUSTER criteria, we again found many situations where the signal was arising because of a few outliers. We decided that what we needed to do was look at clustering the residuals after removing some of the outliers.

One type of outlier in our data is in the X variable (age or, equivalently, date). We sometimes have a single observation (or a small handful of observations) separated from the rest of the observations for the patient by a long period of time – possibly years. This could be valid data – the individual has returned to the study after a long absence – but is more likely an error. The error could be in the date or it could well be an error in the ID variable. Either way, we did not have the resources to study each questionable value. We were looking for situations where there were a substantial number of values that appeared to belong to a different patient. We wanted to set aside the outliers in the X variable so we could focus on the rest of the data.

There are a number of criteria for defining an outlier in a single predictor variable. The standard method for detecting multidimensional outliers in the predictor space is to evaluate the leverage as measured by the diagonal element of the "hat matrix," the matrix $H=X(X'X)^{-1}X'$. The matrix H when multiplied by the vector of observed data produces the predicted values: $\hat{y}=Hy$. The hat matrix is a projection matrix, which means $H^2=H$ and the diagonal elements are no larger than 1. Also, the sum of the diagonal elements is equal to p, the number of predictors in X (including the constant). Belsley, Kuh, and Welsch (1980) provide additional details and suggest the criterion that values greater than 2p/n (twice the average) should be considered as potential X-space outliers. With a single predictor, the mathematical representation of this criterion can be simplified, but since it is easy to get the leverage from PROC REG it seemed worthwhile to maintain the generality in case we use multiple predictor variables later (e.g. adding sex or race/ethnicity to the prediction).

We implemented a simple algorithm for removing potential X value outliers. We used PROC REG to obtain the leverage values H for each data point and then removed the largest one if it exceeded 4/n, where n is the number of data points for that patient. We then repeated the process up to two more times, thereby removing up to three extreme values in the X variable (age).

The next step was to look at outliers in the Y values, the heights themselves. Using the remaining data values (with up to three X outliers removed), we tested for and removed up to three Y outliers. Probably the best indication of an outlier in the Y direction is RSTUDENT, the Studentized deleted residual, which is available from PROC REG. This is the residual you would obtain by fitting the model without that observation divided by its estimated standard error. Values greater than 2 in absolute value are worthy of attention with values greater than 4 being especially concerning. We followed a similar approach as with outliers in the X values, removing up to three Y value outliers sequentially.

We used this approach to remove up to three X outliers and up to three Y outliers, which is reasonable given the amount of data we had available. From the remaining data we again applied PROC CLUSTER to try to identify situations where the residuals from a simple regression of height on age formed two or more distinct clusters. And this produced better results. At last we were finding that many of the patients identified appeared to be strong candidates for having a systematic ID error.

WHAT IS NEXT?

This work is ongoing (we have a lot of data to look at). We are currently looking at using a simple smoothing function to find occasional outliers. Using moving averages omitting the central value, we can look locally for unusual values. Alternatively, it may be useful to use PROC TRANSREG to fit spline functions or other smooth curves, or use PROC LOESS to apply a nonlinear smoother, or use an approach with moving medians (rather than moving averages). Another refinement would be to apply PROC CLUSTER iteratively, so that after identifying one group of observations as potentially belonging to another patient, the algorithm is applied recursively to the remaining data.

CONCLUSION

With enormous amounts of very messy data, we needed to use statistical methods to find the "low-hanging fruit" that represented likely errors in the ID variable. Various approaches we tried resulted in many false positive results – indications of possible ID problems that were in fact much more likely to be errors of other types. We used outlier detection methods to reduce the data for individual patients to avoid those false positives. After eliminating up to three outliers in the X variable (age) and then up to three outliers in the Y variable (height), we found that PROC CLUSTER diagnostics provided a useful metric for identifying possible ID errors in our data. With some variations to suit the specifics of the situation, this approach could be used in many situations where a smooth pattern of data is expected and deviations from that pattern are an indication of possible ID errors. Finally, it is good to remember that graphing the data will almost certainly be helpful.

REFERENCES

Belsley, David A, Kuh, Edwin, and Welsch, Roy E (1980). Regression Diagnostics: Identifying Influential Data and Sources of Collinearity, New York: John Wiley & Sons, 1980.

SAS Institute Inc., SAS/STAT® User's Guide, Version 8, Cary, NC: SAS Institute Inc., 1999.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

David J. Pasta
VP, Statistics & Data Operations
ICON Clinical Research
188 Embarcadero, Suite 200
San Francisco, CA 94105
(415) 371-2111
dpasta@ovation.org
www.iconclinical.com

Eric Elkin
Senior Research Manager
ICON Clinical Research
188 Embarcadero, Suite 200
San Francisco, CA 94105
(415) 371-2153
eelkin@ovation.org
www.iconclinical.com

Rory Michaelis
Statistical Programmer
ICON Clinical Research
188 Embarcadero, Suite 200
San Francisco, CA 94105
(415) 371-2122
rmichaelis@ovation.org
www.iconclinical.com