CENTRACARE Health

Central MN Health Network (CMHN) Health Risk Assessment Instrument – Technology Breakthrough

February 2018

Project Sponsors:

George Morris, MD – Vice President: Performance Excellence
Beth Honkomp , MBA, MSN, RN, NEA-BC – Vice President: Performance Excellence

Project Leader:

Rachael Lesch, RN, BSN, MBA – Director: Clinic Quality Improvement

Project Team:

Thomas Arnold, MS – Population Health Data Analyst: Clinic Quality Improvement

Objectives:

1. Provide an overview of the technology breakthrough with the Central MN Health Network (CMHN) Health Risk Assessment Instrument.

Overview

Risk assessment has important applications in healthcare. The level of available healthcare resources is not sufficient to handle demand. Therefore, individuals must be ranked with respect to their need for healthcare so that preventive efforts can be focused on the patients with the highest needs. Healthcare risk assessment instruments provide a method for creating a rank ordered list of patients in terms of their anticipated need for care.

The accuracy of healthcare risk assessment instruments is assessed using a variety of measures. One of the more important accuracy ratings is the R Squared coefficient, which provides the amount of variation in the outcome variable that is predicted by the Risk Score. Top commercial healthcare risk instruments have an R Square value of about 25%. Previously, CentraCare Staff had been able to develop a Health Spend Risk tool with a 26% R Squared.

By using a new type of mathematical transformation, CentraCare staff have been able to develop a healthcare risk assessment instrument with a 39% R Squared rating. This 50% improvement in R Squared level would appear to be a technological breakthrough that no one else in the healthcare assessment field has discovered. The following document provides a description of the methods used to achieve this level of predictive accuracy.

The tools used in the process would conceptually be categorized as “Population Physics.” The principles of physics and the study of large collections of objects will be applied in the study of human populations. In particular, the work that is being done involves the CentraCare patient population.

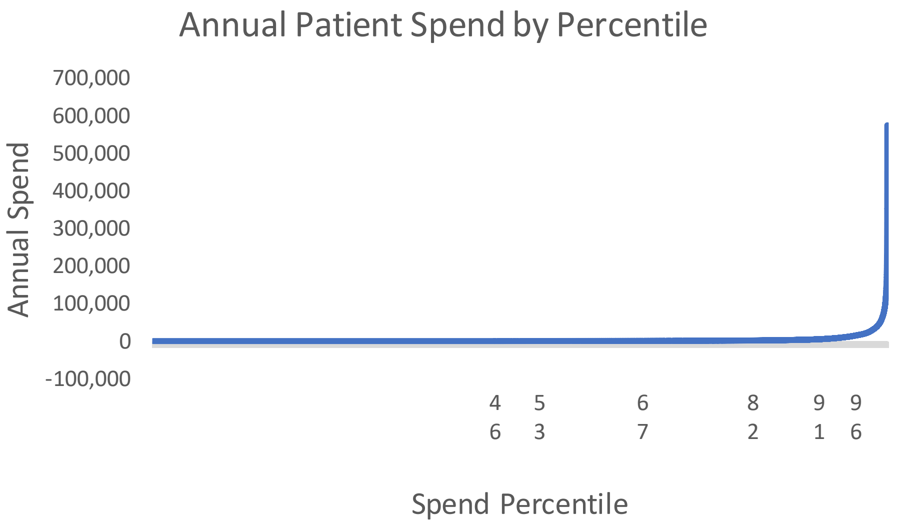
The Problem

The problem with predicting healthcare is that the outcome variable is not linear. Almost half of prior patients do not show up in the healthcare system in each year and the spending is highly skewed towards the highest risk patients. The chart in Figure 1 below illustrates this. This data was taken from CentraCare health spending from 10/1/2016 to 9/30/2017. The first 46% of the patients who had been seen in CentraCare locations in the three years prior to 10/1/2016 had zero spending in the next year. The top 10% of patients had a large portion of the spend, with the highest cost patients costing several hundred thousand dollars to treat.

The nonlinear nature of healthcare spending creates challenges for the creation of healthcare risk scores. The nature of these challenges is discussed at length by Jones, A.M. (2010), “Models for Healthcare.” https://www.york.ac.uk/media/economics/documents/herc/wp/10_01.pdf). The easiest method for creating health risk scores is to use linear regression, but linear regression is not optimal because the outcome variable violates two of the assumptions of linear regression. The outcome is 1) not linear and 2) not normal. However, no one seems to have discovered a way to fix this problem, and most statisticians revert to the use of linear regression in risk score creation. The net result is that risk score accuracies tend to top out at around 25% R Squared.

The technological breakthrough that will be presented in this document is a method to fix the non-linearity problem in healthcare prediction. By using population physics, it is possible to create outcome variables that are much more linear and much more normal. The result is a 50% increase in predictive accuracy.

Figure 1: Annual Patient Spend Distribution



A Theoretical Solution

The solution to the non-linear/non-normal spend problem begins with a theoretical picture of health from a population physics perspective. The theoretical solution to the basic problem of population physics was developed by Thomas Arnold in the analyses of the propensity for criminal behavior. See the work at The Criminological Puzzle for an overview. http://www.thecriminologicalpuzzle.com/ The basic ideas of population physics are applicable to almost any human characteristic. The essential principle is that human conditions are Massively Multivariate, Complex, and Normal (MMCaN).

1. Massively Multivariate: There are an almost infinite number of reasons for human outcomes like healthcare spending or criminal behavior.
2. Complex: The reasons for human outcomes are complex and constantly changing over time.
3. Normal: Because health is a function of Massively Multivariate and Complex factors, the distribution of human characteristics like health is normally distributed.

This theory might not be completely original, but the theory has certain mathematical considerations that are not well understood. That is the discovery that lead to the breakthrough in HealthCare Risk Prediction accuracy.

The MMCaN Model

1: Health Risk is a Massively Multivariate Phenomenon

There are many factors that influence health. A short list would include,

- Genetic
- Developmental
- Situational
- Historical
- Social

These categories include many separate variables which interact with other variables to create an infinite number of causative factors. While some variables have more of an impact than others, the net result is that health caused by an almost infinite number of contributing factors.

2: Health risk is complex, meaning it is highly dynamic in both random and systematic ways

Human beings are complex adaptive systems that are developing over time. In shorter periods, human characteristics are chaotic in nature and fluctuating randomly. Over the life course, human growth and development produces systematic changes in human characteristics. The following provides a discussion of these two types of complex dynamics.

In shorter periods, human characteristics are constantly changing. Change occurs at all time frames, from seconds, to minutes, hours, months, years, and decades. The changes in human characteristics have what is called, “self similarity.” That is, the dynamics of human characteristics are similar for multiple time scales.

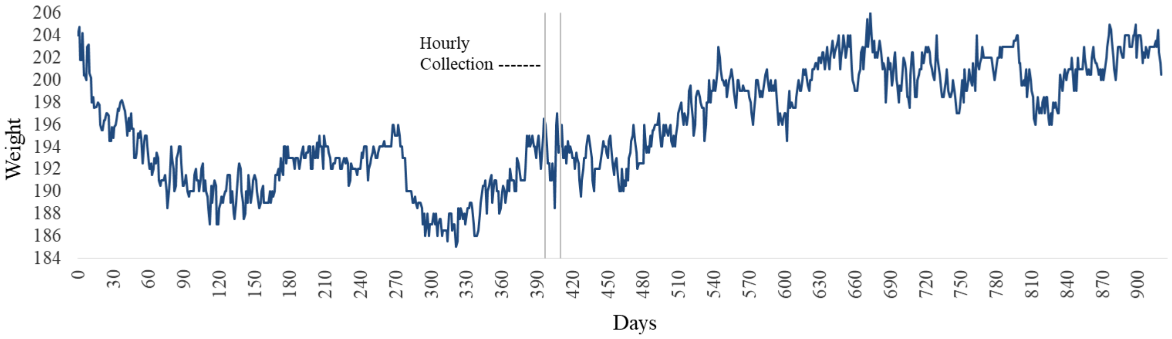
The charts below provide a general sample of how human characteristics change over various time periods. These plots show how my weight changed over different time periods. Although health is much harder to measure than weight, we can assume that health has similar dynamics fluctuation over time. On the flip side, weight affects health, and one can use these charts to see how one health risk factor changes over time.

Dynamic Fluctuation in Shorter Periods Demonstrating Self-Similarity

The first plot in Figure 2 provides a daily weight chart for three years. Note that there is fluctuation in short, medium, and long periods. This is the concept of self-similarity, and is a sign of the chaotic dynamics that underlie human characteristics. There are many variables that affect weight. Exercise, food consumption, motivation levels, current weight, etc. As these various factors interact, weight is constantly fluctuating over time at all time scales.

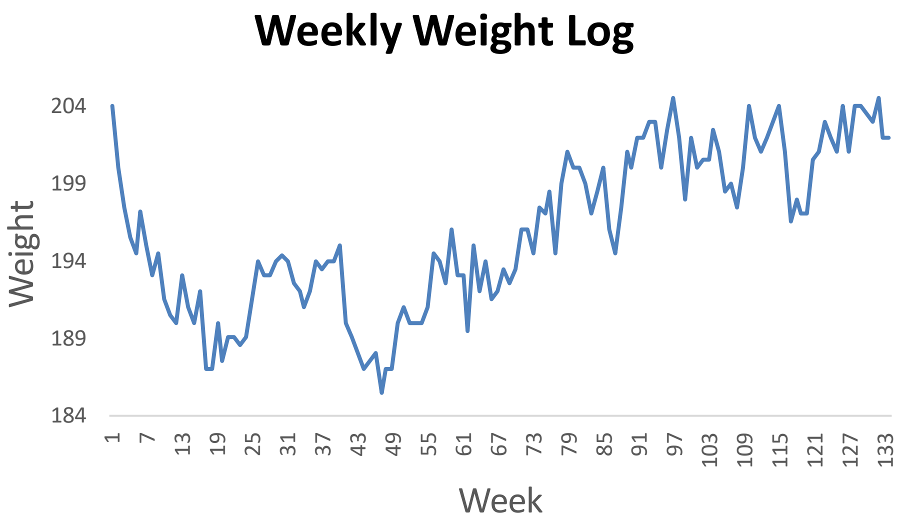
These concepts are directly transferrable to health. Although we cannot measure health as easily as we can measure weight, we know that similar fluctuations must be occurring in health over time.

Figure 2: Three Year Daily Weight Log



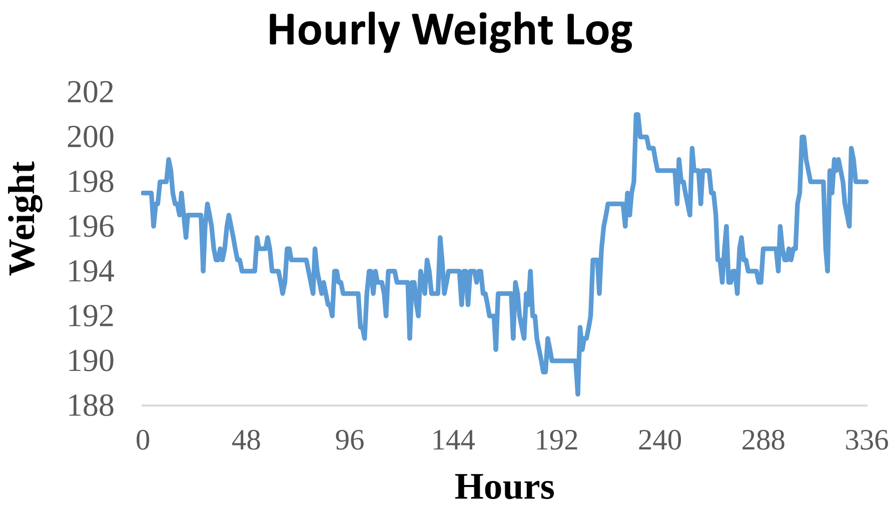
The next chart limits the weight collection period to weeks rather than days. The same pattern of fluctuation is present at this scale. One can seem that if the weight were measured monthly, the same type of fluctuations would be present.

Figure 3: Three Year Weekly Weight Log



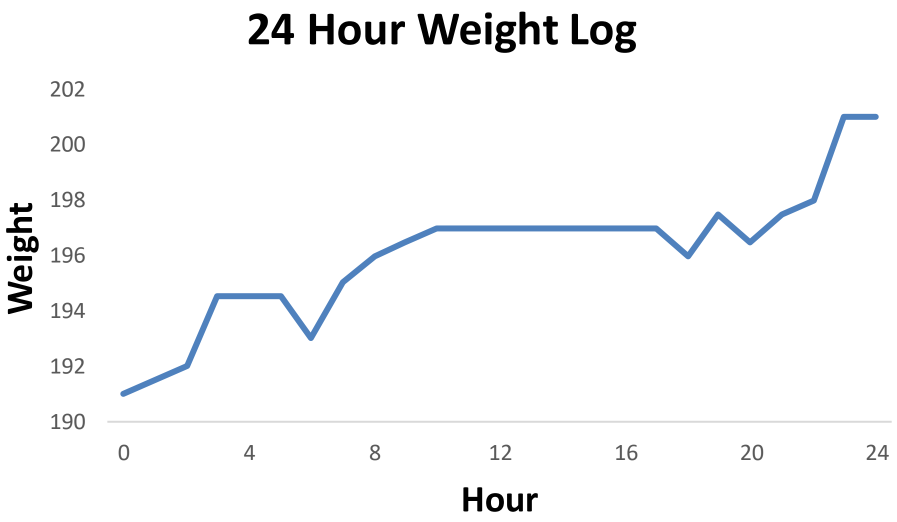
The next chart provides a log of weight collected at hourly intervals. Note that weight was essentially static overnight. The same pattern of fluctuation is present in the hourly measures as in the daily measures.

Figure 4: Two Week Hourly Weight Log



The final chart provides a log of weight collected at hourly intervals over a 24-hour period. Again, weight was essentially static overnight. The same pattern of fluctuation is present in the hourly measures as in the daily measures. Note also that it is probably not a good idea to use one's own measurements to prove a point about weight fluctuation. I gained 10 pounds in one day because I was on vacation and eating whatever I wanted. This is not recommended.

Figure 5: Twenty-Four Hour Weight Log



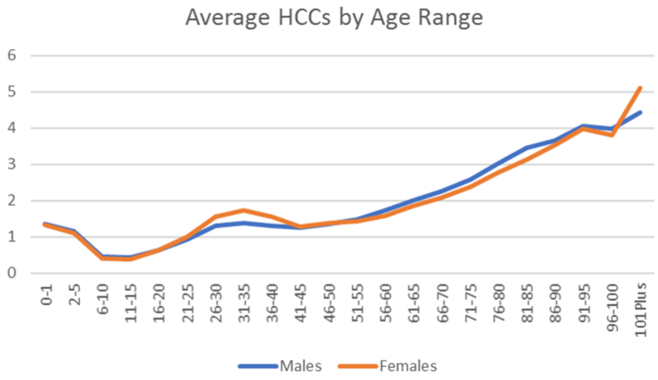
Health risk changes systematically over the life course

Another factor to consider in the complexity of health risk is that health is changing systematically over the life course. If we plot the total number of chronic healthcare conditions using Hierarchical Condition Categories (HCCs) by age, as in Figure 6 below, we find that there is a higher level of risk for infants, health risk is low for toddlers and young children, and then health risk rises steadily over the life course.

Note that females have a slightly higher number of HCCs during the child bearing years, which is not necessarily a measure of poor health, but since having children is a predictor of healthcare spending, and pregnancy is included in the HCC lists.

The fact that health is declining over time is important from a predictive standpoint. Because health risks rise over the life course, it is important to include age as a factor in healthcare risk assessment. We can predict a certain probability of spending simply by knowing a person's age.

Figure 6: Average HCCs by Age Range



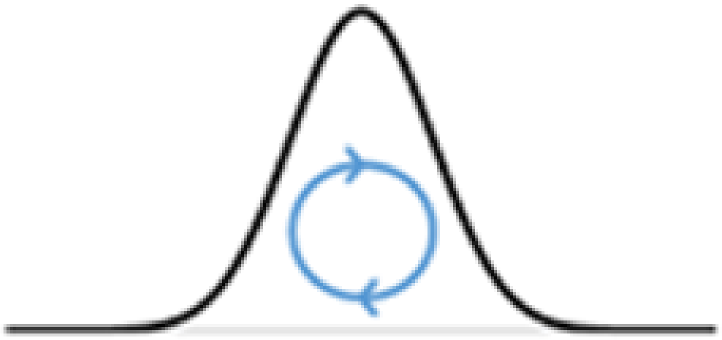
3: Population health risk is normally distributed with mixing

The first two sections demonstrated that health is Massively Multivariate and Complex (MMC). The next step is to demonstrate that health must be normal. Mathematically, this is an almost forgone conclusion because of the central limit theorem. The central limit theorem states that whenever you combine the effects of many independent random variables, the result must be a normal distribution.

The distribution of health risk is not immediately apparent. We really can't measure health risk exactly, and our measures of health risk, like healthcare spending or death rates, are not evenly distributed in the population. However, it can be assumed that health risk is normally distributed in populations. This is a statistical property that flows from the central limit theorem and the joint probability when large numbers of complex factors interact.

Since we know that health risk is fluctuating, we can assume that the normal distribution of health risk is mixing internally. That is, people at the low and high ends of the health risk distribution can be moving to the middle, and people at the middle can be moving up or down. This can be visualized by the following model. The overall distribution of health is stable, even though the health levels of individual people are moving around inside.

Figure 7: A Population Health Risk Model



Methodological Considerations: Health is Massively Multivariate, Complex, and Normal (MMCaN): So What?

The first step in developing a population physics for human conditions was to establish the basic theoretical structure of human characteristics. It can be stated with a high degree of certainty that almost all human characteristics are Massively Multivariate, Complex, and Normal (MMCaN). The pioneering work of Quetelet in the 1800s was instrumental in demonstrating this fact. Further work has corroborated his findings. The next step is to take this information and use it in practical ways.

Probably the biggest shortcoming in the field of population physics has been the lack of understanding of the principles of normal distributions, and this has resulted in a shortage of methods for working with normal population characteristics. We use correlation and linear regression, which presume normal distributions, but people generally do not understand the nature of the variables being used and the nature of the underlying characteristics being analyzed.

The biggest problem is that people generally do not understand the relationship between probabilities and rates. Both probabilities and rates are expressed as percentages, but they are completely different mathematically. A probability is a point estimate, and a rate is the cumulative result of probable events over some period of time. In mathematical terms, the distribution of probabilities can be expressed as a Probability Distribution Function (PDF), and a rate can be expressed as a Cumulative Distribution Function (CDF).

The relationship between the normal PDF and normal CDF is highly nonlinear. The normal PDF is the derivative of the normal CDF, and the normal CDF is the integral of the normal PDF. The underlying formula for the normal distribution is an exponential function with a negative squared value in the exponent. There is a highly nonlinear relationship between normal probabilities (PDFs) and rates (CDFs) that must be understood.

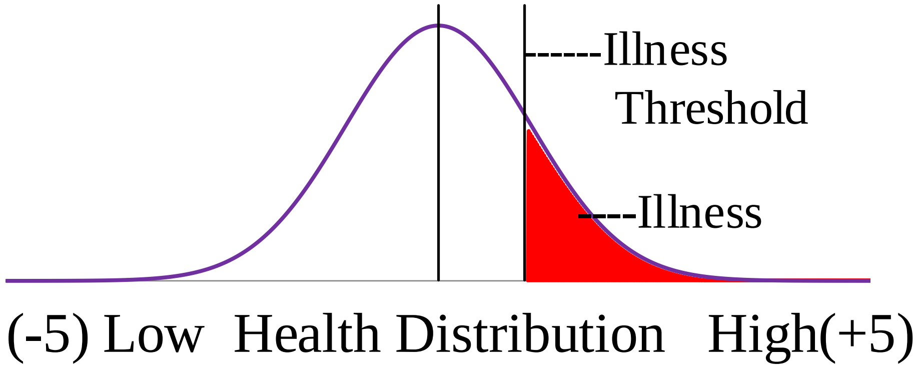
Once the differences between probabilities and rates are understood, this information can be used to enhance risk prediction in healthcare. The following discussion covers the basic mathematical relationship between normal probabilities and normal rates.

Cumulative health risk is an area under a curve

The relationship between health and health outcomes is probably difficult to imagine, since people tend to think in a linear fashion, but health outcomes tend to have a highly nonlinear relationship to health levels. People tend to spend money on healthcare when the illness or condition exceeds a certain level, and not before. Therefore, a threshold model is needed to model health risk. People to the right of the threshold shown below will be likely to spend money on healthcare, while people to the left will not. As people move farther to the right, the amount they will spend increases dramatically.

This model has some interesting implications. One implication is that slight changes in the level of health can have large impacts on participation levels. Conversely, sometimes, large changes in health levels may have little or no effect on participation levels. The reason that there is such a nonlinear relationship between changes in health level and the numbers of people at each level occurs because participation levels are assessed my measuring the area under the curve to the right of the normal distribution. Consider that a one-unit change in the x-Axis under the bulge in the distribution affects a much larger percentage of the population (tens of thousands), and a one-unit change on the edge of the distribution affects only a small percentage of the population (tens).

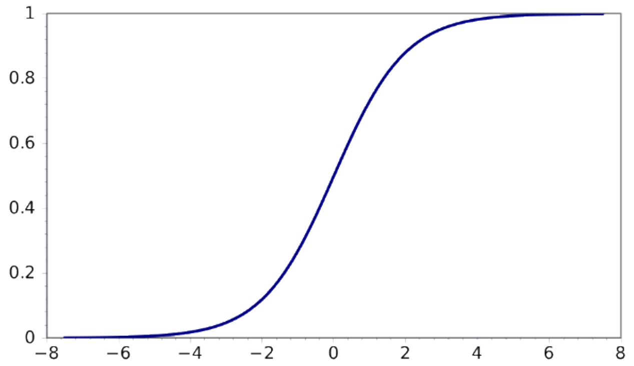
Figure 8: Measuring the Area Under the Curve



A change in mean health risk levels produces a nonlinear sigmoid "S-Shaped" response in participation

Because participation rates in healthcare are assessed my measuring the area under the probability curve (PDF), the rates of healthcare participation for the Cumulative Distribution Function (CDF) follow a nonlinear sigmoid, or "S-Shaped," response curve. For those who have studied dose-response relationships, this will probably seem familiar. The initial effects on an organism of low doses of a poison are small. However, as the dose goes up, the effect rises rapidly until it reaches a plateau. We never reach 100% participation. There are always small numbers of people who can handle elevated levels of a poison with small effect. The response curve illustration the relationship between the health risk PDF and healthcare participation CDF looks like the one shown below.

Figure 9: The Dose-Response Model of Healthcare Risk Vs. Healthcare Participation

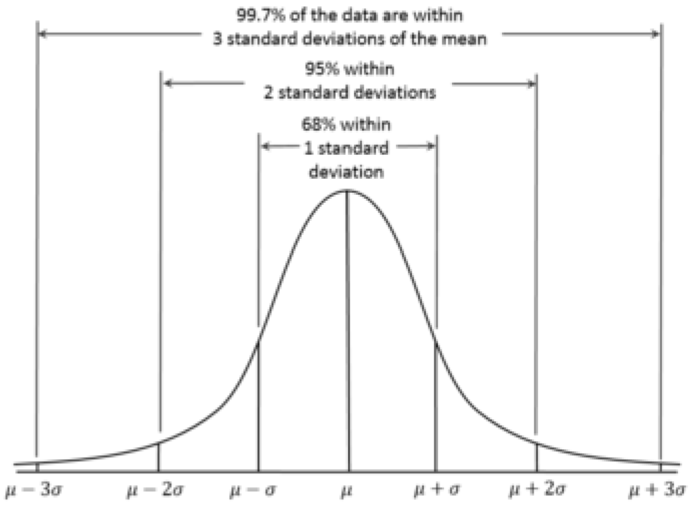


Using Dose Response Models (Probits)

Once we understand the nonlinear relationship between the health risk PDF and the health participation rate CDF, we can use the properties of normal distributions to help us translate between PDF measures (probabilities) and CDF measures (rates). The PDF-CDF transformations were originally developed for assessing the effects of antiseptics on germs, but they work just as well for determining the effects of health risk on health outcomes for populations of people. The basic essence of the dose-response methods is that we can very precisely translate between PDF values (percentiles) and CDF values (Cumulative percentages).

The normal distribution has some consistent properties that can be used to do translation between PDFs and CDFs. It may help to observe the normal distribution with mean μ and standard deviation σ shown below in Figure 10. With a normal distribution, we can convert between percentile ranks and health levels with almost perfect accuracy. Health level would be on the x-axis below and measured in standard deviations from the mean. Because of the properties of normal distributions, we know that exactly .3% of the population is over three standard deviations from the mean health level. Therefore, if we count downward from the top of the healthcare spending table, when we get to the point where exactly .3% of the population is higher spend, we know we have reached the point where patients are exactly 3 standard deviations from the mean health level. We can also do this patient by patient. For example, Patient 1 on the spend list is at 100* 1/318,000 percent and 4.76 standard deviations from the mean to the unhealthy side, Patient 2 is at 100 * 2/318,000 percent and 4.75 standard deviations from the mean, etc. all the way to the last set of patients, who are .23 standard deviations from the mean to the healthy side.

Figure 10: The Normal Distribution Model from Wikipedia



The Creation of Probits

The problem with the previous model is that standard deviations are both positive and negative. This is messy to work with, especially when graphing results, and so Bliss (1935) proposed the "Method of Probits" to overcome this obstacle The method of probits is based on the presumption that normal PDFs almost never see standard deviations more than plus or minus 5 standard deviations from the mean. A 5-standard deviation probability is .000003%, which is almost zero. Therefore, Bliss proposed adding 5 to the normal standard deviation scores to create a "probit" standard deviation scale that varies from 0 to 10 probits, vs. -5 to 5 standard deviations. A linear transformation like this does not affect the distance between the standard deviations, and simply shifts the mean of a standard normal distribution from 0 to 5. We will use the method of probits for plotting outcomes.

Proving Healthcare is Normally Distributed

We have covered the theory of healthcare distributions, the methods for converting probabilities to rates and rates to probabilities, and have a "method of probits" to make the job of plotting outcomes a little simpler. The next step is to prove that healthcare is normally distributed. It is not clear that this has ever been done. I have never seen it, but that is not proof that someone has not done this before.

Theoretically, I assumed that the distribution of health, i.e. the Probability Density Function (PDF) for health is a normal distribution. Therefore, in theory, the rate of participation in the HealthCare system should be the Cumulative Distribution Function (CDF) for a normal PDF. This means that the rate of participation is a sigmoid curve when plotted against the probability of participation. I had no data to support this however, and so it was simply theory.

I used a technique for population physics that I will call "data pattern analysis," to prove that healthcare risk is normally distributed. Essentially, I matched the frequency distribution for chronic conditions to a normal CDF and found almost perfect correspondence. It is proposed that this proves health is normally distributed.

The First Clue – Data Pattern Analysis

This first step simply demonstrated that health is normally distributed. I took a variable measure of poor health (the number of chronic conditions), matched the cumulative frequency to the frequency expected if health was normally distributed, and found almost perfect correspondence.

Epic tracks the number of Chronic Conditions for about 50 different chronic health conditions such as Congestive Heart Failure (CHF), Chronic Obstructive Pulmonary Disease (COPD), etc. The number of the chronic conditions for each CentraCare patients on 10/1/2016 was calculated. Then, the number of patients with each number of chronic conditions was calculated. The results are shown below. About 10.4% of patients had one chronic condition, 6.3% had two chronic conditions, etc.

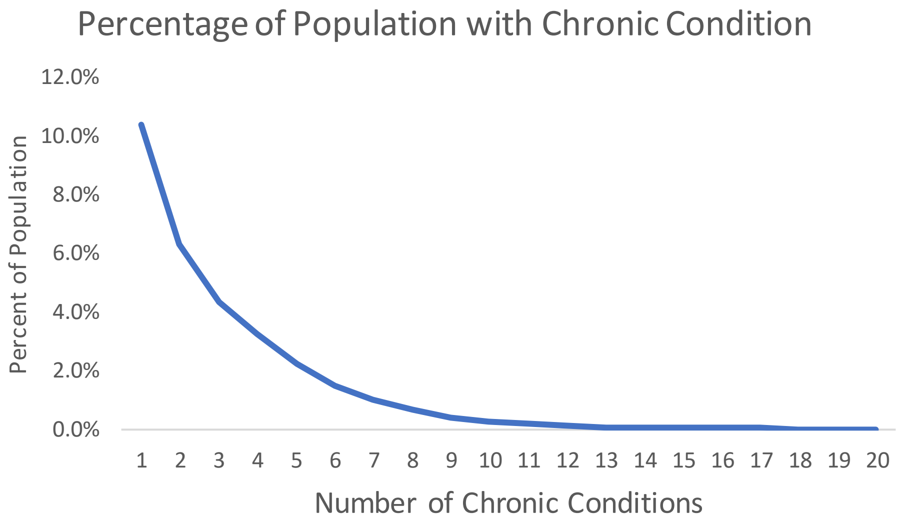
A cumulative percentage was then calculated by adding together all the patients with increasing numbers of chronic conditions. For example, 30.6% of patients had one or more chronic conditions, 20.1% had two or more chronic conditions, etc.

Table 1: Chronic Conditions for CentraCare Patients

Chronic	Number	% With	Cumulative	Cumulative		
Conditions	@ Each	Condition	With 1 or More	%	Z Score	Probit
0	220,702	69.4%	2 or more, etc,			
1	33,059	10.4%	97089	30.6%	-.509	4.491
2	19,963	6.3%	64030	20.1%	-.836	4.164
3	13,656	4.3%	44067	13.9%	-1.086	3.914
4	10,218	3.2%	30411	9.6%	-1.306	3.694
5	7,008	2.2%	20193	6.4%	-1.526	3.474
6	4,762	1.5%	13185	4.1%	-1.734	3.266
7	3,179	1.0%	8423	2.7%	-1.935	3.065
8	2,067	0.7%	5244	1.7%	-2.132	2.868
9	1,288	0.4%	3177	1.0%	-2.326	2.674
10	773	0.2%	1889	0.6%	-2.515	2.485
11	484	0.2%	1116	0.4%	-2.696	2.304
12	288	0.1%	632	0.2%	-2.880	2.120
13	163	0.1%	344	0.1%	-3.067	1.933
14	82	0.0%	181	0.1%	-3.254	1.746
15	59	0.0%	99	0.0%	-3.421	1.579
16	17	0.0%	40	0.0%	-3.660	1.340
17	16	0.0%	23	0.0%	-3.800	1.200
18	3	0.0%	7	0.0%	-4.085	.915
19	3	0.0%	4	0.0%	-4.213	.787
21	1	0.0%	1	0.0%	-4.516	.484
Total	317,791					

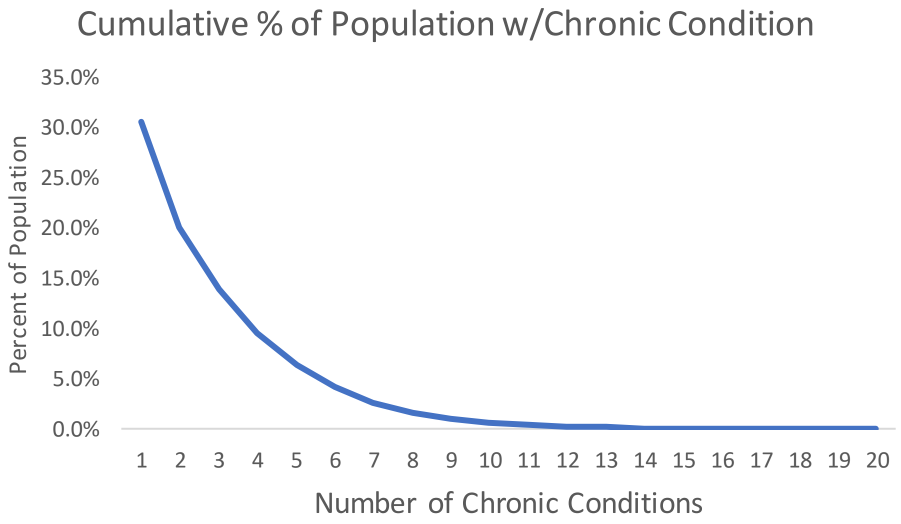
The totals for these various percentages were plotted to get a sense of the distribution shapes. For example, in Figure 11, the number of patients at each chronic condition level were calculated. Attempts to fit a mathematical model to this plot by adding a trendline in Excel were not successful. This frequency distribution did not fit the standard types.

Figure 11: Percentage of CentraCare Patient Population with Chronic Conditions



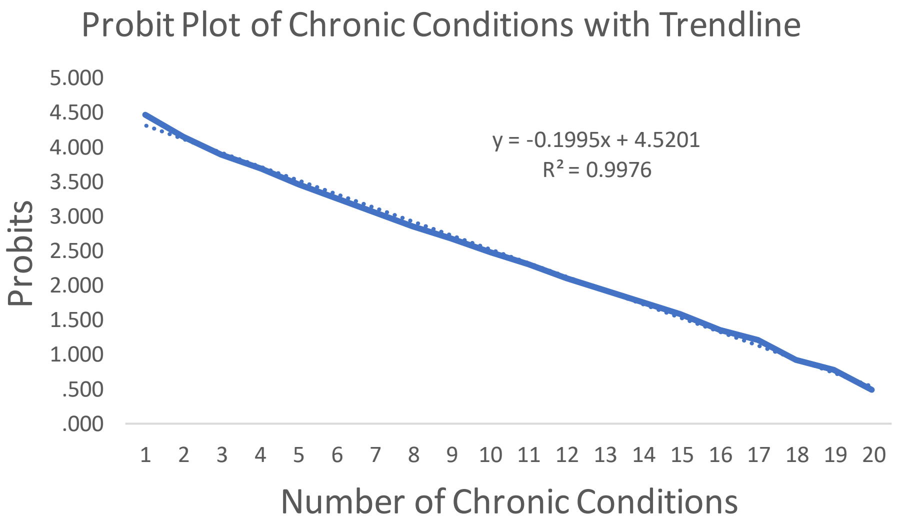
The next step was to plot the cumulative totals. Again, this plot does not fit a standard trendline type in Excel.

Figure 12: Cumulative Percentage of CentraCare Patient Population with Chronic Conditions



Given that the plots shown above in Figures 11 and 12 were not fitting standard plot types, the next step was to try the Normal Distribution model that theory suggests should be generating this data. The first step is to calculate the Z Score for each cumulative percentage. Since the Z Scores are negative, the next step is to apply a probit transformation, which is simply to add 5 to each Z Score. This produced the totals shown on the right columns in Table 1. When the probit values were plotted, the plot was almost perfectly linear. The R Squared for a linear probit trendline was 99.76%, which is almost perfect. This suggests that health is almost certainly normally distributed and that the number of Chronic Conditions is an excellent predictor of health levels for the patients with the most health problems.

Figure 13: Probit Plot of Chronic Conditions with Trendline



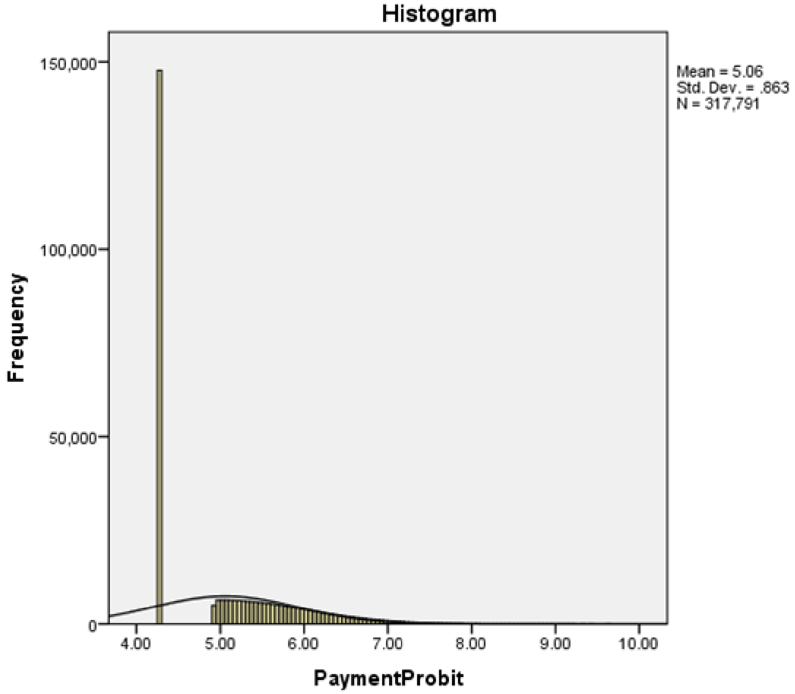
Step Two: Using the Theory to Improve Prediction

Once it was clear that health was normally distributed, the next step was to use this information to try to improve prediction. The Central Minnesota Health Network (CMHN) risk score had been developed previously. The CMHN health risk prediction model uses a linear regression technique to create a health risk score that predicted future spending with an R Squared of about 26%. There were about 480 variables used in the model.

The outcome variable was highly nonlinear, as shown in Figure 1. However, if health is normally distributed, it should be possible to convert the spending distribution to a probit distribution. One simply needed to calculate the standard normal Z score for each patient and add 5. This was done in SPSS using the Rank Cases function on the Transformation menu and selecting Normal Scores with the Blom transformation. The result was called the Payment Probit, and the resulting distribution histogram is shown below.

Note that the PaymentProbit is approximately normal with a huge spike on the left. The spike was caused by the 46% of patients with zero spend in the measurement year. Apparently, the conversion split the 46%, and placed all of the patients at a percentile rank of 23%. One might not expect that this transformation is a big improvement, but the PaymentProbit appears to be a much better outcome variable than the original payment amounts.

Figure 14: PaymentProbit Distribution Histogram with Normal Distribution Trendline



Model Outcome: R Squared = 39.4%

The 480 predictor variables that were used in the CMHN risk model, plus the chronic condition categories from Epic were used in the PaymentProbit model. These included a mix of prior spend, demographics, diagnoses, and chronic condition categories. The data used in this test was captured from Epic with a test start date of 10/1/2016. When the variables were regressed against the PaymentAmount variable, the linear regression model produced an R Squared of 25.3%. This is slightly less than earlier tests with data through 7/1/2016, but similar in magnitude. When the variables in the model were regressed against the PaymentProbit variable, the resulting R Squared was 38.4%. This represents about a 50% increase in predictive accuracy, which is a very sizable improvement. Note that this is different than the 39.4% claimed above, the reason for this discrepancy is explained below.

Converting Back to Something Meaningful

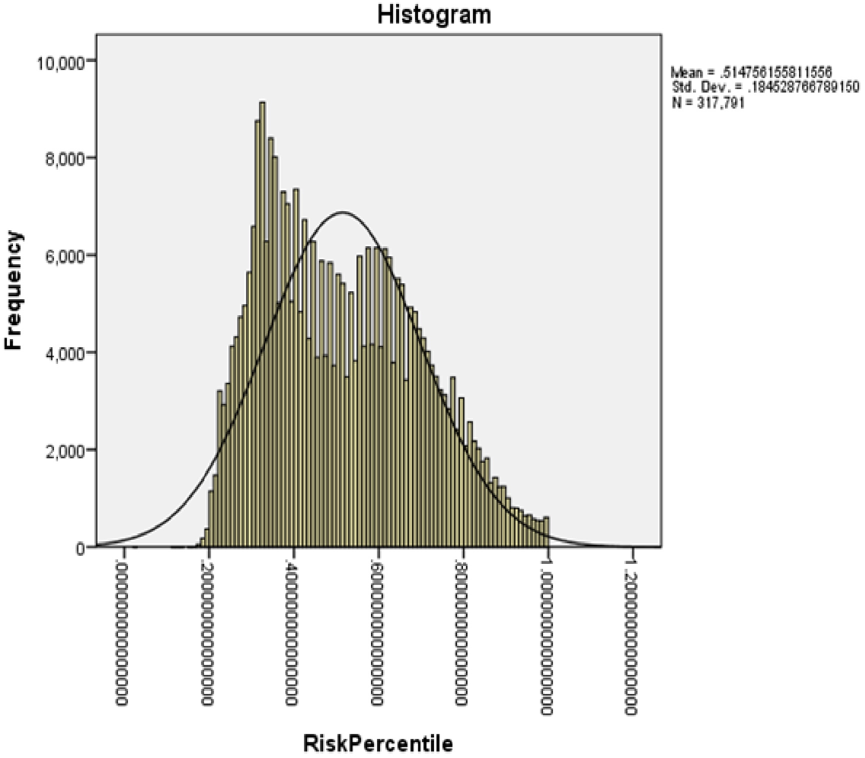
A common problem with nonlinear transformations in linear regression models is that they often produce an output that is hard to interpret. What does a ProbitRiskScore mean? We typically are not used to thinking in Probits or Standard Deviations. It seemed clear that the ProbitRiskScore value needed to be converted to something more meaningful.

Initial attempts to convert the ProbitRiskScore back to an estimate of future payment amounts were not successful. When a lookup table was used to convert Probits to Payment Amounts, the R Squared between the resulting PaymentRiskScore and the Actual Payments dropped to around 11%. This was clearly not an improvement over previous models.

A more acceptable solution was found. Since the output from the ProbitPayment model was a ProbitRiskScore, and Probits are easily transformed back to Z Scores by subtracting 5 from the Probit value, and Z Scores can be used to calculate percentiles, a PercentileRiskScore was created by using a reverse transformation from ProbitRiskScores to PercentileRiskScores. It turned out that the PercentileRiskScore was an even better predictor than the ProbitRiskScore. Note that the independent variable for a PercentileRiskScore is a PaymentPercentile. When predicting the relationship between the PercentileRiskScore and the Payment Percentile for a patient, the R Squared value jumped to 39.4%.

The distribution for the RiskPercentile is somewhat bi-modal in nature. It appears that the patients with no payments and the patients with payments are each having an effect on the distribution of risk scores. This is shown in Figure 15 below.

Figure 15: PercentileRisk Score Distribution Histogram with Normal Distribution Trendline



Conclusion

To my knowledge, no one has ever done this type of probit transformation before with Health Risk. I have worked on the Population Physics model for years and have not found any references to these ideas in any literature. The concepts seem particularly hard for people to grasp. Witness that Quetelet developed some of the basic concepts of Population Physics 180 years ago, and few people have followed his lead. There is something called Probit Regression and it is used in certain situations, but I do not see it used like this. Therefore, this appears to be a completely new discovery. It seems certain that if people had figured this out, the predictive accuracies in commercial risk predictors would be higher.

The essence of this probit transformation method is as follows.

1. Convert Payment Amounts to Probits.
2. Regress the Payment Probits against the standard risk predictors to create a Probit Risk model.
3. Calculate a ProbitRiskScore from the Probit Risk model.
4. Convert the ProbitRiskScore to a PercentileRiskScore.

This method appears to increase the predictive accuracy of the CMHN risk model by 50%. This may not be all that is possible. This method might also be used on the predictor variables to improve model accuracy even further. For example, the spending in previous years is also non-linear. Probit transformations on these predictor variables could improve predictive accuracy even further. Tests will have to be done to explore the limits of this model.

This method seems valid. The data produces similar magnitude R Squared values with the base CMHN model, so there does not appear to be a mistake in the collection process. Since all that was done was to do some simple transformations, the transformations appear to have created the improvement in predictive accuracy.

There does not seem to be a real problem with converting from a Payment Prediction model to a Payment Percentile model. We were converting risk scores to percentiles before. This eliminates a step. The rank order accuracy is higher, so it should help produce a better tool for patient management.

Moving forward, some thought might be given to trying to capitalize on this technological breakthrough. It is not clear whether this idea could be patented. Probit models have been used for a long time. It may be worth shopping this idea to some companies such as Milliman or Optum that use healthcare risk prediction methods. However, it is not clear how one could protect the intellectual capital once it becomes more widely known.