Is Whole Milk Good For You?

Denny Anderson
University of Virginia
dra2zp@virginia.edu

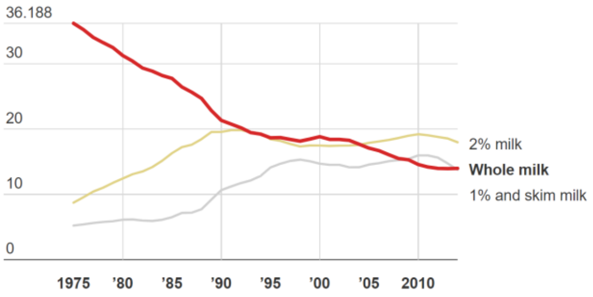
1 ABSTRACT

The goal of this paper is to analyze the parameters and factors that were involved in the statistical model that the government used to show that saturated fats caused heart disease. This caused the government to issue a blanket statement that said saturated fat consumption was bad and that whole milk should be avoided as it contains a lot of saturated fat. This paper also considers improvements to statistical models (namely, predictive models) that could have been used instead to avoid this erroneous conclusion. This paper argues that statistical models are not equipped to handle large data with several different features since those models favor simplicity over accuracy. Predictive models are much better because they are more accurate at the expense of being more complex. But for an issue that has the potential to alter the eating habits of the population, it is important to obtain the most accurate data as possible.

2 INTRODUCTION

In the United States before the 1970s, the consumption of whole milk was widely encouraged. People of all ages were expected to drink at least a couple glasses of milk each day to maintain healthy bones, and children and teenagers were told they would grow up healthy and strong. Whole milk contained several vitamins and nutrients that people required, including carbohydrates, protein, vitamin A, calcium, and vitamin D.

However, starting in the 1970s, the government and public health officials recommended that consumers drink fat-free (skim) milk because whole milk was high in saturated fats, which led to an increased risk of heart disease, according to statisticians. As a result, public whole milk consumption severely declined over the period from 1975 to 2015, and the consumption of fat-free and reduced-fat milk increased over the same period (Figure 1) [4].



Source: USDA

The Washington Post

Figure 1 – Milk consumption in the United States from 1975 to 2015.

This shows that the government's recommendation on what the population should and should not be consuming has a strong impact on people's food choices. When public health officials determine that it's unhealthy to continue consuming something, people tend to listen. People usually consider the government's advice as fact rather than merely a possible suggestion made by a statistical model. Thus, it is of utmost importance that the government has access to precise and unflawed data, so that their statistical models can be as accurate as possible.

In the case of whole milk, the government was provided a statistical model that suggested a correlation between saturated fat intake and an increased risk of heart disease. This paper attempts to analyze the factors that went into this faulty statistical model leading to an incorrect blanket assumption that whole milk is bad due to its saturated fats, as well as consider possible alternatives in the field of predictive modeling in order to better model and understand the data before any blanket conclusion is made by the government.

3 STATISTICAL MODELING

Statistical methods are used to understand how the input variables explain some output [2]. They attempt to understand the relationship between input and output variables by estimating a probabilistic model from observed data [2]. Usually, many models are fit in search for a 'simple' model involving a small subset of input variables [2] . Regression is the most common form of a statistical probabilistic model with a linear regression being the simplest.

Occam's Razor is used to justify choosing a simpler model because a simple model is more easily understood. However, in prediction, accuracy and simplicity seem to be at odds with each other. Linear regression gives a highly understandable picture of the relationship between *x* and *y*, but its accuracy is usually less than that of neural nets, which can be highly complex [1]. Accuracy generally requires more complex prediction methods [1]. Simple and interpretable functions do not make the most accurate predictors.

The factors leading to the government's erroneous conclusion that whole milk is bad most likely comes from a faulty statistical model that sacrificed accuracy for simplicity in order for the model to be more interpretable. This stems from Breiman's notion of the two-culture system: the data modeling culture, which the vast majority of statisticians fall under, and the algorithmic culture, which contains very few statisticians and more people from other fields like machine learning [1]. This two-culture system and statistician's tendency to focus on data modeling rather than algorithmic modeling, according to Breiman, has led to irrelevant theory and questionable scientific conclusions, has kept statisticians from using more suitable algorithmic models, and has prevented statisticians from working on exciting new problems [1].

This was certainly the case for whole milk. Statisticians most likely used data models instead of algorithmic models, and as a result, an incorrect conclusion was made about the healthiness of whole milk.

Next, we look at predictive models and contrast them with statistical models to find out if the whole milk controversy would have been different had a predictive model been used instead.

4 PREDICTIVE MODELING

Unlike statistical data models, predictive models are not concerned with simplicity. Instead, they are focused on being as accurate as possible. Breiman shows through several examples that random forests are capable of discovering important aspects of the data that standard data models cannot uncover [1]. This is because random forests are much too complex for statisticians to consider when working with a data model, where simplicity is chosen over accuracy.

Breiman also states that higher predictive accuracy is associated with more reliable information about the underlying data mechanism [1]. Weak predictive accuracy can lead to questionable conclusions [1]. Algorithmic models can give better predictive accuracy than data models and provide better information about the underlying mechanism [1].

Data in the medical field is highly complex, mostly because our bodies are highly complex. There are too many features (parameters) to consider when trying to create a model. It is very likely that when statisticians created a data model, they only focused on a very few number of features in order for their model to be as simple as possible. Their model probably only had two variables: saturated fat intake and heart disease. While this is a perfectly fine model to consider, it certainly should not be the only model that is used to form a conclusion. This is where predictive modeling is used to form a more accurate, albeit more complex, model for the data.

Breiman suggests that data scientists should be open to using a wide variety of tools [1]. This means that they should combine several of their simple data models to form a more accurate hypothesis regarding the data. Thus, they shouldn't just look at one simple model showing the correlation between saturated fat consumption and risk of heart disease. An important notion that Heid discusses is that correlation is not causation, meaning that simply because a higher saturated fat intake appears is linked with a higher risk of heart disease does *not* mean that saturated fat *causes* heart disease [3]. There could be other features that the data model does not account for that explains the higher risk for heart disease. Statistical models would not show this, but as Breiman said, a good statistician should use a wide variety of tools, and this includes looking at other parameters in relation to heart disease to ensure that saturated fat truly is the cause. Predictive models, especially random forests and decision trees, would have done much better in finding features of interest that are most closely associated with heart disease since the feature would be located at the top of the tree (the head node).

Predictive and algorithmic models would have probably been able to figure out that saturated fats alone do not cause heart disease, especially those found in whole milk. While most saturated fats do indeed raise the levels of “bad” cholesterol found in the body, those found in whole milk have been shown to contribute to the levels of “good” cholesterol [4]. In addition, a study that used more data with predictive models may have shown that by replacing whole milk with alternatives such as bread, cookies, and cakes have been found to increase risk of heart disease [4].

It is also important to note that drinking skim or reduced-fat milk causes people to feel less full than drinking whole milk. This most often causes people to continue eating or drinking until they feel fuller [3]. Also, there have been psychological studies showing that when people consume foods marked as "low fat," they feel as if they can eat more of that food since it contains less fat, and they end up eating more and increasing their fat intake than if they would have just opted for the full-fat version.

Predictive models would have been better equipped to handle more complex parameters than statistical models, and they most likely would have shown the truth about whole milk rather than the ultra-simplistic hypothesis that saturated fat from whole milk causes heart disease, as the data model showed.

5 CONCLUSION

In conclusion, statistical data models should never be used alone. They only use a select subset of all the possible features available, which can cause erroneous conclusions to be made. They also favor simplicity over accuracy, and since they use the fewest number of features possible in the model, it can seem like correlation implies causation when that is certainly not the case. Statisticians must use all their available resources and consider multiple regression algorithms and parameters before making any sort of conclusion about the given data. Predictive models are generally considered more favorable since they are far more accurate than statistical models, and there are techniques like decision trees that can be used to find a better fit for the data. The government's use of a statistical model that caused a general statement about the unhealthiness of saturated fats in whole milk to be issued illustrates just one example of the importance of more accurate models. Public health officials should have as much accurate information as possible before giving advice to the population about what they should or should not consume. Our health is highly important, so more accurate predictive algorithmic models as opposed to statistical data models in the case of whole milk should have been used before issuing any sort of statement affecting our health.

REFERENCES

[1] Breiman, Leo. "Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author)." *Statistical Science* 16, no. 3 (2001): 199-231. Accessed November 04, 2017. https://projecteuclid.org/Euclid.ss/1009213726.

[2] Cherkassky, Vladimir. "Statistical Methods." In *Predictive Learning*, 161-80. Minneapolis, MN: Vladimir Cherkassky, 2013.

[3] Heid, Markham. "Why Full-Fat Dairy May Be Healthier Than Low-Fat." Time Inc. March 5, 2015. Accessed November 4, 2017. http://time.com/3734033/whole-milk-dairy-fat/.

[4] Whoriskey, Peter. "For decades, the government steered millions away from whole milk. Was that wrong?" The Washington Post. October 06, 2015. Accessed November 04, 2017. https://www.washingtonpost.com/news/wonk/wp/2015/10/06/for-decades-the-government-steered-millions-away-from-whole-milk-was-that-wrong/?utm_term=.6cf9b5fb5dc5.