**Lab 9: DIABETES DATA SCIENCE**

I. Introduction

II. Blood glucose regulation

III. Type 2 diabetes

IV. Introduction to the dataset

\*\*Note: This lab will be conducted entirely within the Jupyter notebook environment, using the lab computers or your laptops. The lab manual here introduces you to the physiology, but the exact directions for the lab will be in the Jupyter notebook, which you will submit to receive credit, similar to a worksheet. \*\*

**I. INTRODUCTION**

When you get a medical examination, various measurements are collected, like your blood pressure, weight and the concentration of various molecules and cells in your blood. This data is used by doctors to diagnose diseases or to act as warning signs of future problems. For example, if blood glucose levels in a fasted patient are 100-125 mg/dL, the patient is considered to be prediabetic and they should take action to lower their glucose levels through lifestyle changes and/or medication. Knowing who is at risk of developing a disease helps doctors prepare the best treatment plans for each patient.

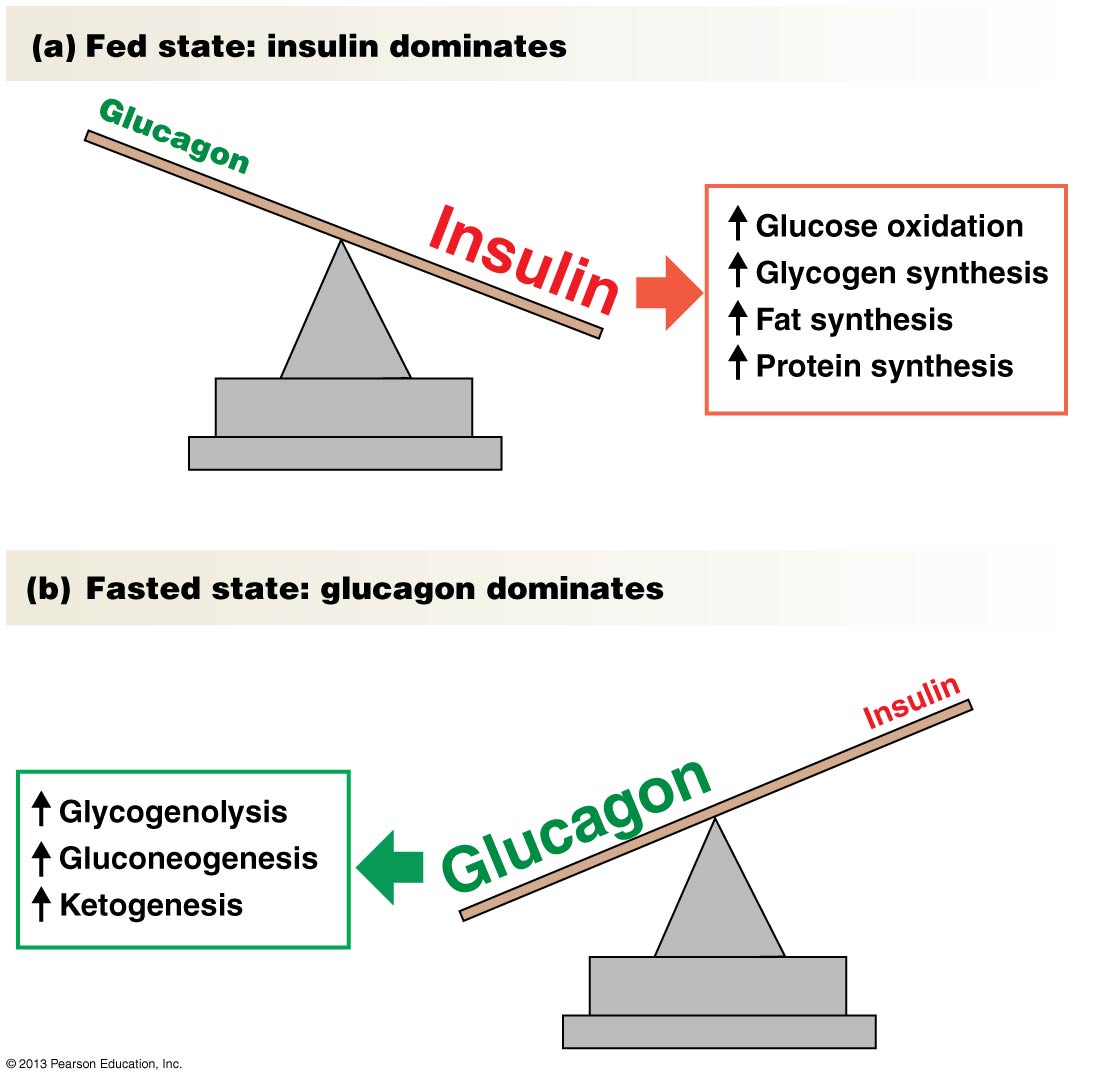
For complex diseases like diabetes or cardiovascular disease, there are many factors that correlate with the disease. Some factors might be causing the disease while others are changes that result from the disease. Looking at large patient datasets with lots of variables can help us learn more about the biology of the disease, as well as allow us to develop models to predict disease risk.

In this lab we will use a dataset of individuals who have type 2 diabetes and compare them to healthy controls. We will look at how well different variables correlate with the disease and we will use machine learning to look for patterns and develop a **classifier**. The classifier is an algorithm that can be used on other data to predict who has diabetes or is at risk of developing it. We can test the classifier to see how accurate it is using another dataset. The Lab 9 Jupyter notebook will guide you through this process. As you work with the data keep in mind the biology of diabetes, which will be introduced in the next sections.

**II. BLOOD GLUCOSE REGULATION**

Glucose is used by all the cells in our body to generate ATP, necessary for enzymatic reactions, membrane transport, muscle contraction, etc (see lab 8). Glucose from food is absorbed into the body in the intestines. Excess glucose is stored as the polysaccharide **glycogen** in skeletal muscles and the liver. If we have not eaten recently (the “fasted” state), then the body signals the liver to breakdown the glycogen and release glucose into the blood. Thus, blood glucose levels need to be highly regulated, so there is always enough energy available to our cells. Blood glucose homeostasis is regulated by two hormones released by the pancreas: insulin and glucagon (Fig 1).

**Insulin** is released by beta cells in the pancreas after meals when blood glucose levels are high. Insulin signals for cells to insert the Glut4 glucose transporter into the plasma membrane, allowing cells to uptake and use glucose from the blood. Insulin also signals the liver and muscles to do glycogenesis, which is the synthesis of glycogen from glucose.



**Fig. 1** Regulation of blood glucose by insulin and glucagon. Note that they regulate both glucose metabolism and fat metabolism. (Figure from Silverthorn *Human Physiology*)

**Glucagon** is released by alpha cells in the pancreas during fasted states when blood glucose levels begin to decrease. Glucagon signals the liver and muscles to do glycogenolysis, or to breakdown glycogen stores. In addition, glucagon signals the synthesis of glucose from amino acids (gluconeogenesis). Both of these processes increase the amount of glucose in the blood, bringing levels back within the normal range of 90-100 mg/dL.

**III. TYPE 2 DIABETES**

The importance of maintaining blood glucose homeostasis is clear when we learn about diabetes, which is a failure of the body to regulate glucose levels. There are two main types of diabetes mellitus: type 1 and type 2 diabetes.

* **Type 1 diabetes** is an autoimmune disorder where the body destroys beta cells and the patient has to inject themselves with insulin to survive.
* **Type 2 diabetes** is the more common type, which often starts when insulin receptors become less sensitive to insulin. As the disease progresses, the beta cells try to compensate by releasing more insulin, but this can damage the beta cells. Insulin release is then impaired and the patient has to take synthetic insulin. Type 2 diabetes is a complex disease caused by both genetics and the environment.

In both types of diabetes, untreated patients have high glucose levels in their blood, because insulin cannot properly signal cells to take up the glucose. Without that glucose, cells will not be able to make as much ATP, which can cause the diabetic patient to feel tired and weak. In addition, high blood glucose itself can impair various physiological processes like wound healing and kidney function, as well as damage small blood vessels and neurons. Without proper treatment, diabetes is a progressive and deadly disease. Luckily, there are some treatments available. Patients with type 1 diabetes can monitor their blood glucose levels and inject themselves with the appropriate amount of insulin. For patients with type 2 diabetes, the main treatment involves losing weight, getting exercise and taking medications to boost insulin signaling.

As of 2015, 9.5% of Americans had been diagnosed with type 2 diabetes, and it is the 7th leading cause of death in the United States. This is an important public health issue and yet it is still somewhat of a mystery why some people get diabetes and others do not. In order to understand the causes of diabetes, the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) have conducted a number of studies of diabetic patients and have made some of that data available to the public. We will be using one of these datasets to learn about classification and to understand more about what causes diabetes.

**IV. INTRODUCTION TO THE DATASET**

We will use a dataset from NIDDK, which collected measurements from female diabetes patients and female controls, both of Pima heritage. The Pima are a group of Native Americans from regions in Arizona. The prevalence of diabetes is much higher in the Pima (~40%) than in other groups living in the U.S. Studies of the Pima in the U.S. have shown that genetic variants and changes in lifestyle have contributed to this high rate of diabetes.

The variables in the dataset include:

* Number of pregnancies
* Glucose: blood glucose levels 2 hours after an oral glucose tolerance test (mg/dL)
* Blood pressure: diastolic blood pressure (mm Hg)
* Skin Thickness: the thickness of a skin fold at the triceps (mm)
* Insulin: insulin levels in the blood 2 hours after oral glucose tolerance test (μU/mL)
* Body mass index: weight (in kg)/(height (in m))2
* Diabetes pedigree function: takes into account how many of the patient’s relatives have diabetes and how many do not, the age at which they were diagnosed and how related they are to the patient (a parent vs a cousin). The greater the value, the greater the “genetic risk” of developing diabetes.
* Age
* Outcome: 0 = no diabetes, 1 = diabetes (as determined by glucose measurements)

Follow the directions in the Lab 9 Jupyter notebook to investigate this dataset and generate classifiers. Submit the notebook to get credit for completing the Lab 9 worksheet.