# Measuring Patient Similarity



## Introduction

In patient care, various categories of illnesses are determined based on symptoms, medical history, laboratory results, and other diagnostic criteria. Just as farmers and scientists categorize new fruits to understand their properties and marketability, healthcare professionals classify illnesses to understand their nature and determine appropriate treatment strategies. Today, your task is to classify a patient's illness based on a set of symptoms, and their medical history.

## **Part I: Patient Similarity**

Below is a table of symptoms and medical histories for four patients.

| **Patient** | **Fever** | **Cough** | **Stuffy Nose** | **Drug allergy** |
| --- | --- | --- | --- | --- |
| Rue | Yes | No | Yes | No |
| Devi | Yes | Yes | Yes | No |
| Serena | No | Yes | No | No |
| Belly | Yes | No | No | Yes |

1. Which two patients are most similar? Explain.
2. Come up with and describe a way to **quantify** **how similar** two patients are based on their symptoms and medical histories.
3. Use your method to compute how similar Devi and Belly are.

## Binary Attributes and Simple Matching Coefficient (SMC)

A binary attribute has only two possible outcomes. One example of a binary attribute would be whether a person has a college degree or not; the only outcomes are ‘yes’ or ‘no’. In our patient data, all of the attributes are binary.

Because of the categorical nature of our attributes, the two distance measures we have learned about thus far (Euclidean distance and cosine similarity) are not useful—how do you calculate the difference between “Yes” and “No”? One simple metric for computing the similarity between two sets of categorical attributes is the *Simple Matching Coefficient* (SMC). The SMC is defined as the proportion of elements that match between two sets of categorical attributes. For example, consider the set of symptoms for Devi and Belly:

| **Patient** | **Fever** | **Cough** | **Stuffy Nose** | **Drug allergy** |
| --- | --- | --- | --- | --- |
| Devi | Yes | Yes | Yes | No |
| Belly | Yes | No | No | Yes |

There is one symptom on which Devi and Belly match (fever) out of four total symptoms being considered. Thus the SMC is ¼ = 0.25. Note that although it did not appear in this example, a symptom that neither Devi nor Belly had (both were “No”) would also constitute a match. A formula for the SMC is:

Where nYES,YES is the number of symptoms that both patients have a value of “Yes” for, nNO,NO is the number of symptoms that both patients have a value of “No” for, nYES,NO is the number of symptoms that the first patient has a value of “Yes” and the second patient has a value of “No”, and nNO,YES is the number of symptoms that the first patient has a value of “No” and the second patient has a value of “Yes”. In our example,

Note the numerator is simply summing the number of matches, and the denominator is the total number of symptoms being compared.

1. Compute the SMC between Devi and Serena.
2. What if two patients matched on every single symptom. What would the value of the SMC be for these patients?
3. What if two patients had no symptoms in common. What would the value of the SMC be for these patients?
4. Based on the SMC, is Devi more similar to Belly or Serena? Explain?
5. Compute the SMC between each set of two patients.
6. Which two patients are most similar? Which are most different?

## Symmetric versus Asymmetric Binary Attributes

When we use the simple matching coefficient we typically make the assumption that the binary attributes are *symmetric.* A symmetric binary attribute is one in which both outcomes (e.g., Yes and No) are equally important and carry the same weight or significance. In contrast, an *asymmetric* binary attribute is one in which the two outcomes have different levels of importance or significance. For example, consider the potential results of a pregnancy test: positive (indicating pregnancy) or negative (indicating no pregnancy). While the results are binary, they are asymmetric—a positive result typically has bigger implications than a negative result.

It is debatable whether the patients’ symptoms are symmetric. The presence of a symptom is often more significant than the absence of a symptom. If the binary attributes are asymmetric the use of the SMC can be misleading. For asymmetric binary attributes, since the matched presence of the symptom matters more than the matched absence of the symptom—the value of |1–1| should carry more weight than the value |0–0|. One similarity measure that accomplishes this is the *Jaccard Index*. The Jaccard index is computed as:

This measure is computed in a similar manner to the SMC, but in both the numerator and denominator we ignore the number of symptoms that both patients do not have, that is we ignore the matches deemed less important in the asymmetry. FOr example, to compute the Jaccard index between Devi and Belly, we use:

In this example, the value of the SMC and Jaccard index are the same. That is not always the case.

1. Compute the Jaccard index between each set of two patients.
2. Which two patients are most similar? Which are most different?