A Bayesian approach to chest pain

O. Kastelianou, V. Konstantakos
MSc in Data Science: Large-scale statistical methods
NCSR "Demokritos" & University of the Peloponnese

1 Introduction

Chest pain is a common chief complaint with a wide range of conditions that can cause it, starting from illnesses with favorable prognosis to life-threatening conditions. However, translating the patient's experience of pain in the chest to a specific pathology continues to haunt the workload of clinicians. The consequences of an incorrect diagnosis are, at times, severe; misdiagnosing a heart attack as minor musculoskeletal pain could mean death. Thus, the challenge is to timely identify an acutely dangerous cause, but on the other hand, avoid unnecessary testing and referrals.

Unfortunately, doctors often rely on pattern recognition to arrive at the appropriate diagnosis. This, however, is not always reliable as many cases can initially present in an uncommon way. Clinicians that have been confronted with an unexpected diagnosis of a myocardial infarction in a seemingly innocuous presentation of chest pain can confirm this. A solution to such diagnostic problems is to adopt a more systematic Bayesian approach, incorporating multiple factors before arriving at the final decision.

1.1 Bayes' Theorem

Bayes' theorem is a formula that describes the probability of an event, based on prior knowledge of conditions that might be related to the event. It follows simply from the axioms of conditional probability, but can be used to powerfully reason about a wide range of problems involving belief updates such as medical diagnosis. It is stated mathematically as following:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{1}$$

where A, B are events, $P(A)$, $P(B) \neq 0$ are the probabilities of observing A and B, while $P(A|B)$ and $P(B|A)$ are the conditional probabilities of A given B and B given A, respectively. While this is an equation that applies to any probability distribution over events A and B, it has a particularly nice interpretation in the case where A represents a hypothesis H (or multiple hypotheses H_i) and B represents some observed data D. In this case, the formula - for one or n hypotheses, respectively - can be written as:

$$P(H|D) = \frac{P(D|H)P(H)}{P(D|H)P(H) + P(D|\neg H)P(\neg H)} \quad \text{or} \quad P(H_i|D) = \frac{P(D|H_i)P(H_i)}{\sum\limits_{j=1}^{n} P(D|H_j)P(H_j)} \tag{2}$$

where $P(H_i|D)$ is the posterior probability of hypothesis i, given the data, $P(D|H_i)$ is the likelihood of observing the data under hypothesis i and $P(H_i)$ is the prior probability for hypothesis i. The denominator includes the likelihood of observing the data partitioned for every hypothesis (i.e., 2 or n, respectively) and functions as a normalizing constant.

1.2 Problem formulation

The latter form of Bayes' theorem is especially useful for testing multiple hypotheses simultaneously. We apply this to study the challenges that occur during the differential diagnosis of chest pain. Specifically, we will try to answer the following scenario: "A male patient presents with chest pain as the chief complaint. What is the probability that he has a myocardial infarction? In general, how likely is it to have a cardiovascular, respiratory, etc., disease? Furthermore, how do these diagnoses change for a female patient? What are the corresponding probabilities in that case?" This report will try to tackle these questions in a brief but comprehensible manner and explain the basic concepts underlying this approach.

2 Methods

2.1 Prior distributions

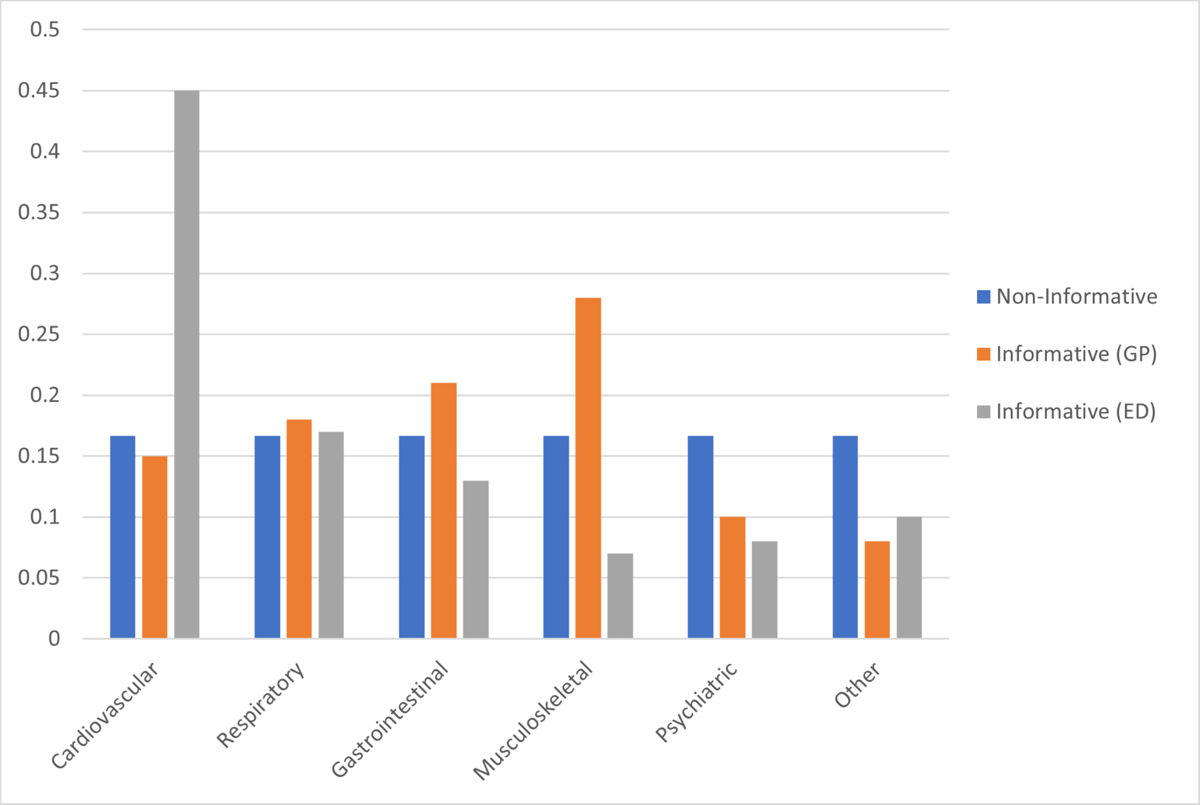
To begin with, we create 30 discrete hypotheses for a patient presenting with chest pain. We assume that these hypotheses are mutually exclusive (i.e. if one is true, the others must be false) and exhaustive (i.e. all hypotheses, taken together, describe all possible outcomes). For example, a patient with chest pain will have one (and only one) diagnosis from those 30, in our schema.

We then construct a prior distribution, considering each hypothesis as equally likely. Thus, each diagnosis will have an equal (1/30) probability and the resulting distribution will be non-informative. To better represent the existing knowledge, we also create 2 informative prior distributions, each one reflecting a different population. In particular, we construct a prior distribution for the patients presenting with chest pain in General Practice (GP) and one for the patients presenting in the Emergency Department (ED). Previous studies have shown that these populations are significantly different and, as such, they should be studied separately. To accomplish this, we create 1000 virtual visits - in GP and ED - based on our intuition and prior knowledge. Each visit has a final diagnosis, which is then used to count and calculate a corresponding probability. For instance, we observed 50 myocardial infarctions among 1000 visits in GP, resulting in a probability of 0.05. Due to space limitations, we group the 30 diagnoses to 6 categories based on the biological system to which they belong (e.g., cardiovascular, respiratory, etc.). A summary of these distributions is shown in Supplementary Tables S1 and S2. The differences between the three prior distributions are clearly demonstrated in Fig. 1. We observe that cardiovascular diseases are the most prevalent among the visits in the ED while musculoskeletal causes are quite common in GP. Furthermore, psychiatric or other miscellaneous causes are, in reality, not that common in contrast to what a non-informative distribution would illustrate. Thus, it is crucial to create and select a suitable prior distribution for the problem we are studying because it will directly influence our results.

2.2 Likelihood calculation

After constructing the appropriate prior distributions, we need to calculate the likelihood of observing the data under each of our hypotheses, as we can see from Eq. 2. This is not always an easy task, especially for numerous discrete hypotheses, or even worse for continuous hypotheses. In our case, we use the data from 22304 visits to separate them into a unique diagnosis and identify how many of those were male patients. We can then compute the likelihood of being male under each possible hypothesis $[P(Male|H_i)]$; dividing the number of male patients by the number of total visits for each diagnosis gives the desired result. In addition, we group the diagnoses into 6 categories, as we did before, for a clearer illustration. An added benefit of that calculation is

Figure 1: Prior distributions for patients presenting with chest pain



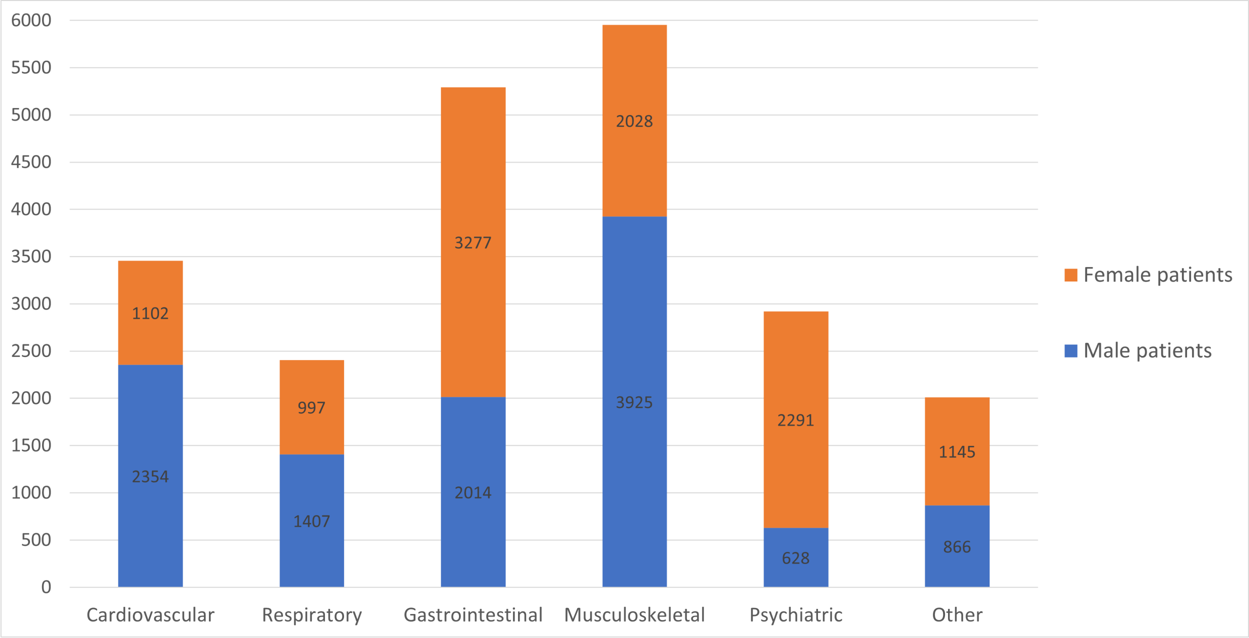
that we can also derive the likelihood of being female under each possible scenario, as those events are complementary (i.e., a patient can be either male or female) (Fig. 2). We use this fact as a final comparison in our analysis in the following section. First, we focus our attention to the initial problem of a male patient presenting with chest pain. All the resulting likelihood calculations can be seen in Tables S3, S4.

We observe that the sum of all likelihoods isn't equal to 1. Is this outcome expected? In addition, could we use the likelihood of each hypothesis divided by the sum of all likelihoods as the new likelihood? These are questions that commonly arise when dealing with such problems and should be considered. We try to answer them after completing our analysis in the following section.

2.3 Posterior Distributions

We now proceed to calculate the posterior distributions based on the computed likelihoods and the prior distributions we have defined. In particular, we use the data from Table S1, multiply them with the corresponding likelihood from Table S4, calculate their sum (i.e., the denominator) and finally arrive at 3 posterior distributions for every diagnosis. Due to space limitations, we will present the results grouped into 6 representative categories for demonstration purposes. The complete results for each diagnosis and category can be seen in Table S5 and S6, respectively.

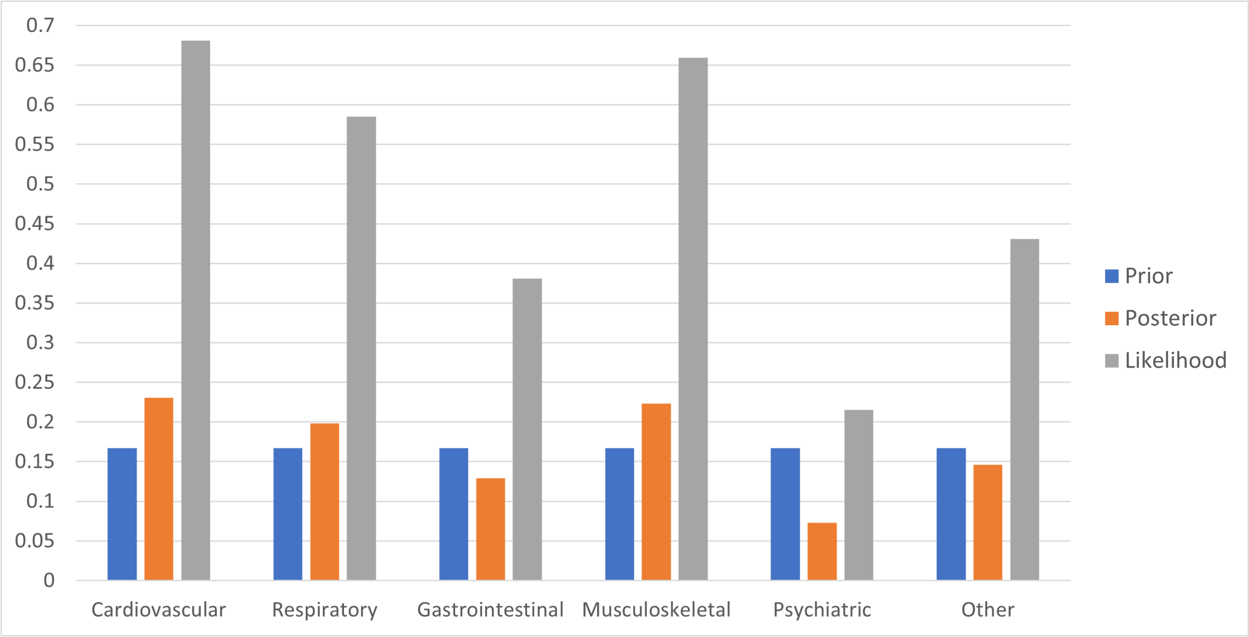
Figure 2: Visits presenting with chest pain per diagnostic category



3 Results

In this section, we present the results of our analysis. To begin with, we illustrate the constructed posterior distributions after following the process we described.

Figure 3: Posterior distribution with non-informative prior



In particular, Fig. 3 shows the posterior distribution that arises from the non-informative prior we used. In this case, the data show that most male patients with chest pain had cardiovascular, musculoskeletal or respiratory disease as their likelihoods were higher. After observing the data, the posterior probabilities for these conditions were increased. Therefore, we observe that the posterior mainly follows the trend of the likelihood; thus, it is data-driven. However, even though there is a strong bias towards cardiovascular and musculoskeletal diseases for male patients the resulting probabilities do not have a substantial difference from the initial ones.

On the other hand, Fig. 4, 5 illustrate a different posterior distribution for each setting. Specifically, the prior probability for GP was dominated by musculoskeletal diseases. Likewise, the data show that most male patients have cardiovascular, respiratory or musculoskeletal diseases (i.e., their likelihoods are higher). After observing the data, the posterior probabilities for those three were increased (more for the latter one). It follows that male patients presenting with chest pain in GP have a vastly increased probability to suffer from a musculoskeletal disease. In contrast, gastrointestinal diseases are pretty common in GP but affect female patients more frequently; thus, the probability of a male patient presenting with chest pain due to such conditions is generally low.

The distribution for the patients in the emergency department (ED) also differs significantly. In particular, chest pain presentations in that healthcare setting are dominated by cardiovascular diseases (Fig. 5). For this reason, male patients - who are already at risk for cardiovascular conditions (i.e., high likelihood) - will very likely face such a diagnosis. In contrast, the increased likelihood of male patients having a musculoskeletal disease only slightly changes its posterior probability. In this case, the shape of the prior has a strong influence and drives the posterior distribution in a major manner.

Figure 4: Posterior distribution with informative prior (GP)

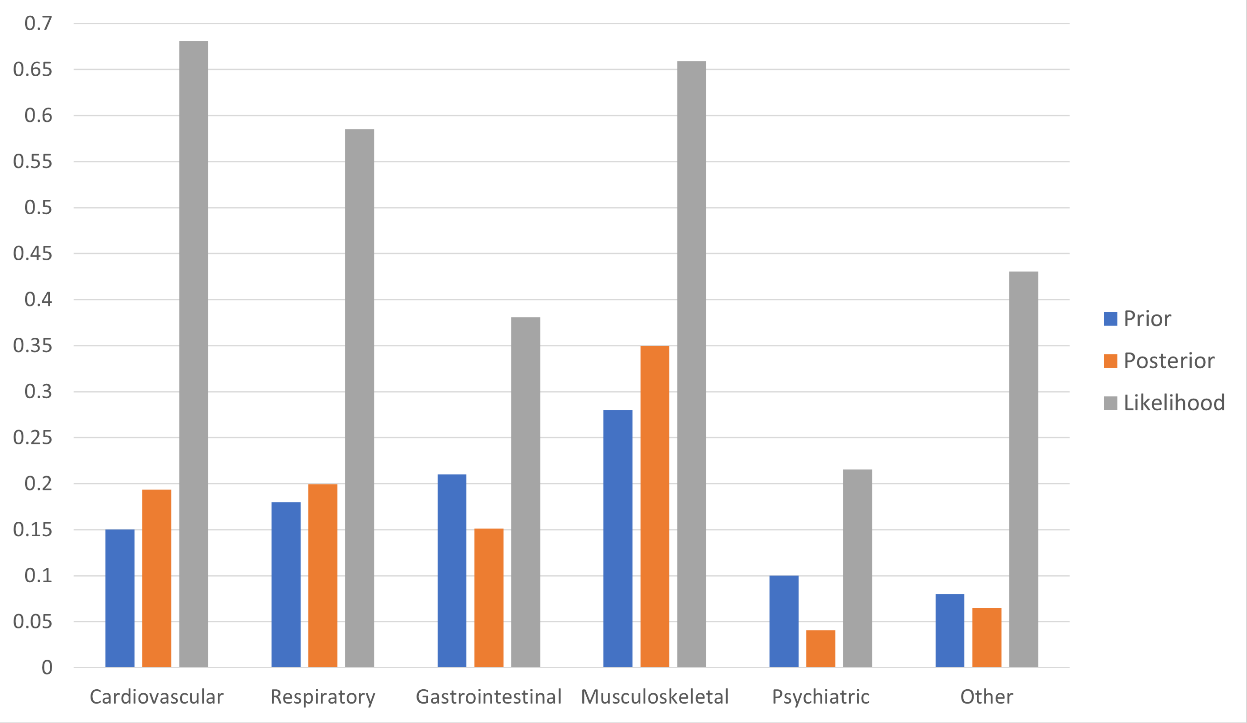


Figure 5: Posterior distribution with informative prior (ED)

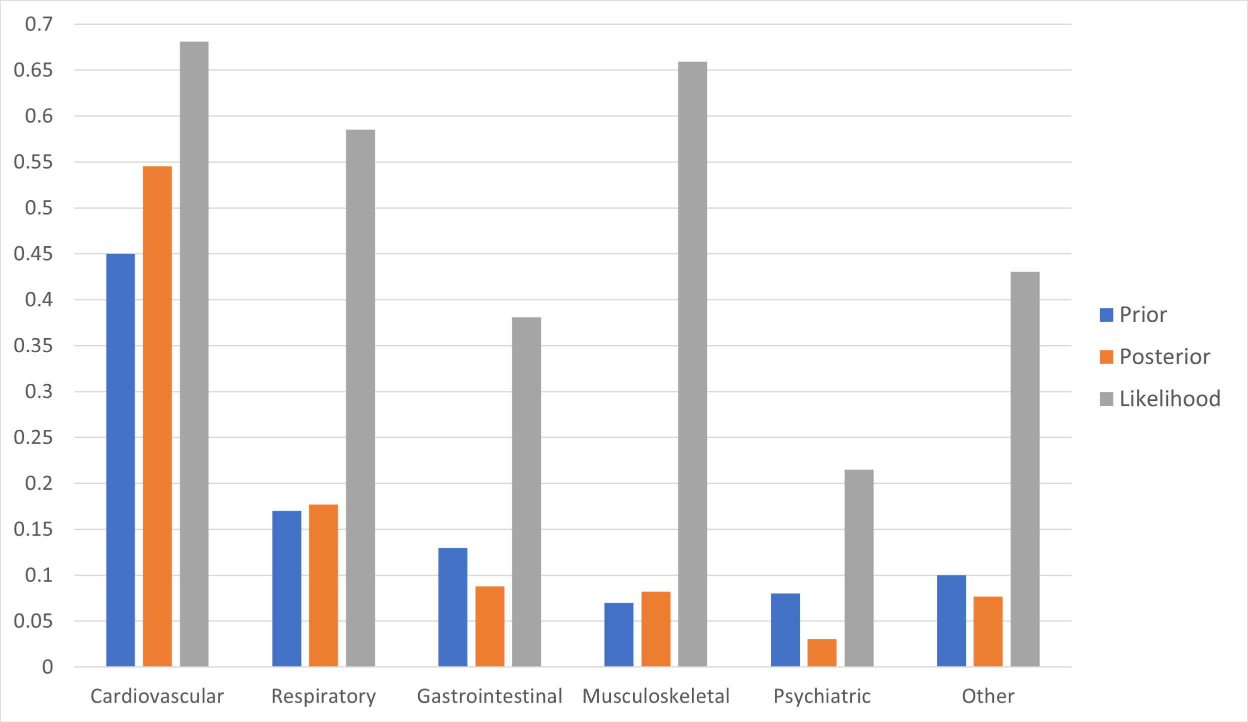


Figure 6: Posterior distributions for male patients presenting with chest pain

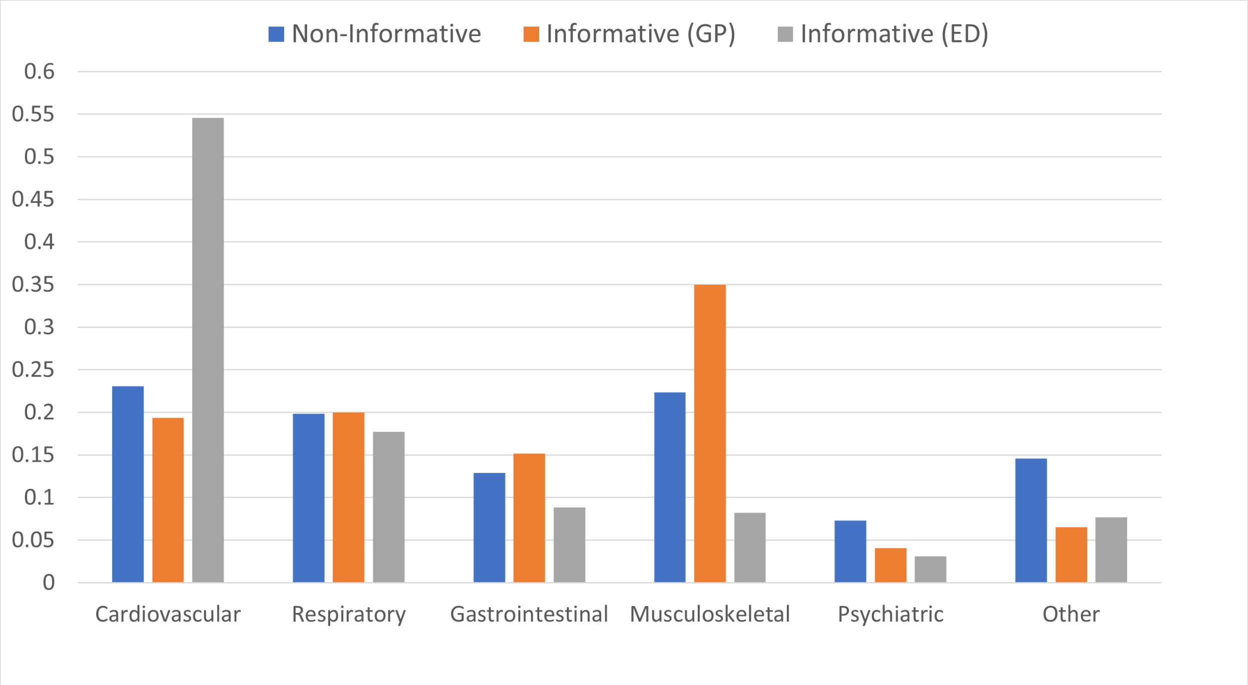
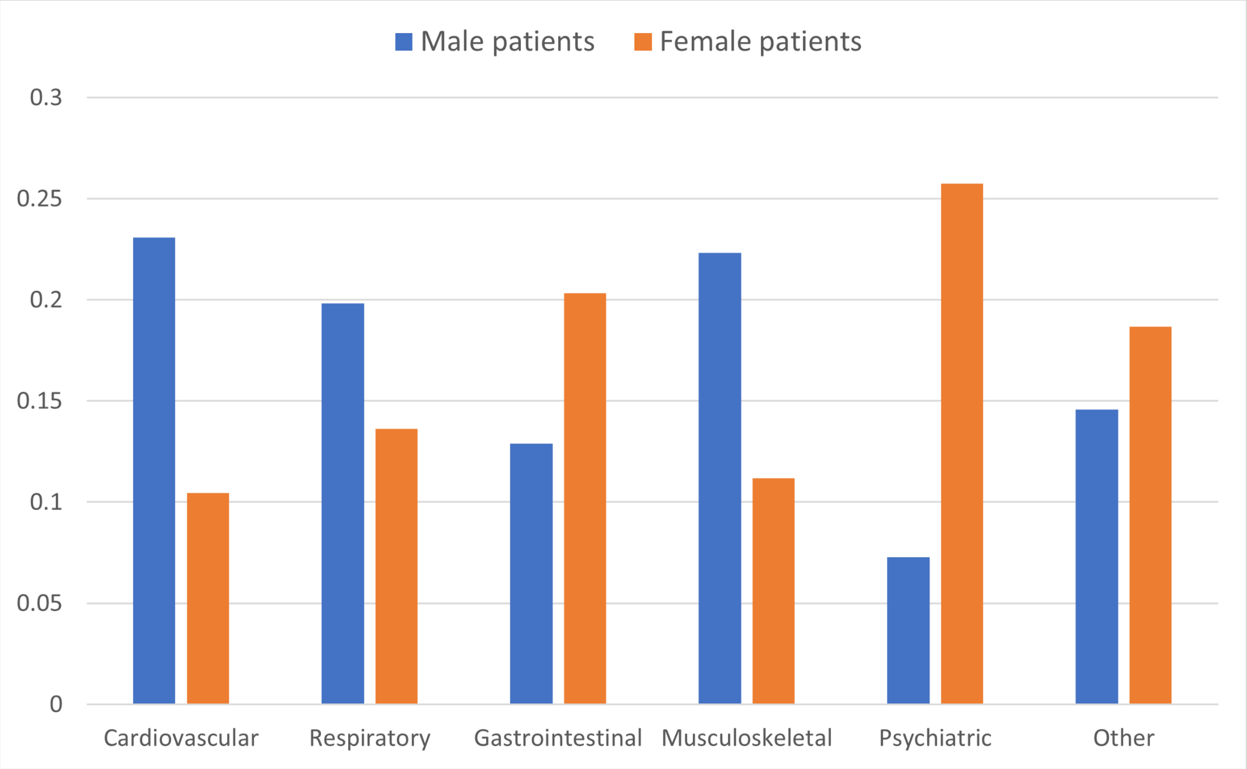


Figure 6 further visualizes the differences between the mentioned posterior distributions. We confirm the importance of both the gender (male patient) and the healthcare setting (GP, ED) when evaluating the posterior distribution. Cardiovascular and musculoskeletal diseases in ED and GP, respectively, should be first considered when a male patient presents with chest pain. On the other hand, psychiatric and other miscellaneous causes of chest pain are fairly uncommon for male patients. Finally, we provide a direct comparison of the posterior distributions between both genders (Fig. 7). We chose to only compare the ones that involve a non-informative prior to better capture the concept of the likelihood shaping the posterior. We notice that female patients suffer more often from gastrointestinal, psychiatric or other diseases when presenting with chest pain. We also confirm that the likelihood primarily drives each posterior distribution. In fact, by comparing Fig. 2 and Fig. 7, it is clear that this is the case, as they demonstrate a quite similar trend.

Figure 7: Posterior distributions for male/female patients (Non-Informative prior)



4 Discussion

To summarize our results, a non-informative prior gives equal probabilities for each hypothesis and allows the posterior to be shaped only by the likelihoods. We demonstrate this in two ways in our analysis. Regarding the two informative prior distributions, we also notice interesting results. In the first case (GP), we observe that male patients have increased probabilities for cardiovascular and musculoskeletal diseases. This is in agreement with both the increased prior and the increased likelihood for male patients. Thus, both the prior and the likelihood contribute to the final outcome. On the other hand, the presentation in the emergency department (ED) provides a unique picture. Specifically, while the likelihood for the three mentioned diseases remains high, the resulting probabilities are significantly different only for the cardiovascular ones. The other two are just slightly increased. Therefore, the prior drives the posterior in a major manner in this healthcare setting. Furthermore, in Section 2.2 we asked the following questions:

- We observe that the sum of all likelihoods isn't equal to 1. Is this outcome expected?
- Could we use the likelihood of each hypothesis divided by the sum of all likelihoods as the new likelihood? Does this ratio define something meaningful?

In fact, this outcome is expected. As we see in Tables S3, S4, the sum of all possible likelihoods (e.g. 2.95, 13.84) is not equal to 1 but the likelihoods for each condition separately add up to 1. These two observations actually answer our question. To be specific, when we are summing the likelihoods of each diagnosis, we are summing probabilities from different probability spaces. Thus, the result doesn't need to be 1; it could be less or greater than 1. Formally, the likelihood $P(D|H_i)$ is not a distribution in H_i, but only in D. In our case, where the data is the gender of the patient - male or not male (i.e., female) - summing across this axis gives the expected result of 1 (Tables S3, S4)

Regarding the second question, there are two important facts we need to note. First, if we divide the likelihood of each hypothesis by the sum of all likelihoods and use the resulting likelihood as the new one, the results will be exactly the same. The only difference would be that the likelihoods will now add up to 1 as we have just performed a normalization procedure. We can formally prove this as follows: Let's denote as $L_i = P(D|H_i)$, $P_i = P(H_i)$ the likelihood and prior probability under each hypothesis H_i, respectively. In addition, we define S as the sum of all likelihoods (i.e. $S = \sum\limits_{i=1}^{n} L_i$) and $L'_i = \frac{L_i}{S}$ as the new likelihood. Then the new posterior probability for each hypothesis H_i is equal to:

$$P'(H_i|D) = \frac{L'_i P_i}{\sum\limits_{j=1}^{n} L'_j P_j} = \frac{\frac{L_i}{S} P_i}{\sum\limits_{j=1}^{n} \frac{L_j}{S} P_j} = \frac{L_i P_i}{\sum\limits_{j=1}^{n} L_j P_j} = P(H_i|D) \quad \text{with} \quad \sum_{i=1}^{n} L'_i = \sum_{i=1}^{n} \frac{L_i}{S} = 1 \tag{3}$$

This answers the first part of the second question. Finally, the ratios $L'_i = \frac{L_i}{S}$ can also define something meaningful. In particular, they represent the posterior probability for each hypothesis when all the hypotheses have the same prior (i.e. when we use a non-informative prior). Indeed, if all the prior probabilities for each hypothesis were equal to some constant c ($P_i = c$) then the posterior for a hypothesis H_i would be:

$$P(H_i|D) = \frac{L_i P_i}{\sum\limits_{j=1}^{n} L_j P_j} = \frac{cL_i}{\sum\limits_{j=1}^{n} cL_j} = \frac{L_i}{\sum\limits_{j=1}^{n} L_j} = \frac{L_i}{S} = L'_i \tag{4}$$

which is the new likelihood that we have defined. Actually, this further illustrates the fact that when we are using a non-informative prior distribution, the likelihood drives the posterior. If we are only making use of the data, this is precisely what we get. One can also confirm this by looking at the "Likelihood Ratios" and "Non-Informative" columns in Tables S3 and S6, respectively.

5 Conclusion

In summary, this case study presents an overview of how Bayes' theorem works and how it can be incorporated into medical diagnosis. Specifically, Bayes' theorem provides an extended, systematic framework to approach complex problems which involve belief updates. Medicine is a domain that involves such problems and demands the integration of multiple factors before reaching the final decision. Therefore, understanding the concepts that underlie Bayesian inference, its benefits and challenges will surely pave the way to its more accurate and widespread application. Pattern recognition can perform well when patients fit the pattern. When they don't, a more formal, systematic approach that is based on probability theory is needed. This is what Bayes' theorem provides.

6 Supplementary Material

Table S1: Prior distributions per diagnosis for patients presenting with chest pain

Diagnosis	Non-Informative	Informative (GP)	Informative (ED)
Myocardial Infarction	0.033333333	0.05	0.17
Angina	0.033333333	0.05	0.13
Pericarditis	0.033333333	0.03	0.08
Myocarditis	0.033333333	0.01	0.05
Aortic dissection	0.033333333	0.01	0.02
Pneumonia	0.033333333	0.05	0.07
Pleuritis	0.033333333	0.02	0.02
Asthma/COPD	0.033333333	0.08	0.04
Pneumothorax	0.033333333	0.02	0.03
Pulmonary embolism	0.033333333	0.01	0.01
Gastroesophageal reflux	0.033333333	0.08	0.01
Peptic ulcer disease	0.033333333	0.07	0.03
Esophageal dismotility	0.033333333	0.03	0.02
Acute cholecystitis	0.033333333	0.02	0.04
Acute pancreatitis	0.033333333	0.01	0.03
Costochondritis	0.033333333	0.11	0.01
Trauma	0.033333333	0.09	0.02
Fibromyalgia	0.033333333	0.03	0.01
Neoplasms	0.033333333	0.03	0.02
Rheumatoid arthritis	0.033333333	0.02	0.01
Panic disorder	0.033333333	0.03	0.03
Generalized anxiety disorder	0.033333333	0.02	0.01
Major depressive disorder	0.033333333	0.02	0.02
Somatic symptom disorder	0.033333333	0.02	0.01
Deception syndromes	0.033333333	0.01	0.01
Herpes zoster	0.033333333	0.02	0.02
Sarcoidosis	0.033333333	0.01	0.01
Substance related	0.033333333	0.03	0.03
Systemic lupus erythematosus	0.033333333	0.01	0.02
Acute chest syndrome	0.033333333	0.01	0.02
Total	1.0	1.0	1.0

GP: General Practice, ED: Emergency Department. Each horizontal line separates a unique diagnostic category.

Table S2: Prior distributions per category for patients presenting with chest pain

Diagnostic Category	Non-Informative	Informative (GP)	Informative (ED)
Cardiovascular	0.166666667	0.15	0.45
Respiratory	0.166666667	0.18	0.17
Gastrointestinal	0.166666667	0.21	0.13
Musculoskeletal	0.166666667	0.28	0.07
Psychiatric	0.166666667	0.1	0.08
Other	0.166666667	0.08	0.1
Total	1.0	1.0	1.0

GP: General Practice, ED: Emergency Department.

Table S3: Likelihood of each gender per diagnostic category

Categories	Likelihood of being male	Likelihood of being female	Total	Likelihood Ratios
Cardiovascular	**0.681134259**	0.318865741	1	0.230724008
Respiratory	**0.585274542**	0.414725458	1	0.198252968
Gastrointestinal	0.380646381	**0.619353619**	1	0.128938249
Musculoskeletal	**0.65933143**	0.34066857	1	0.223338627
Psychiatric	0.215142172	**0.784857828**	1	0.072876182
Other	0.430631527	0.569368473	1	0.145869967
Total	2.9522	3.0478	6	1

Likelihoods ≥ 0.6 are marked in bold.

Table S4: Likelihood of each gender per diagnosis

Diagnosis	Likelihood of being male	Likelihood of being female	Total
Myocardial Infarction	**0.708635997**	0.291364003	1
Angina	**0.696760855**	0.303239145	1
Pericarditis	0.622680412	0.377319588	1
Myocarditis	0.582938389	0.417061611	1
Aortic dissection	0.571428571	0.428571429	1
Pneumonia	0.532314924	0.467685076	1
Pleuritis	0.5275	0.4725	1
Asthma/COPD	0.640664962	0.359335038	1
Pneumothorax	**0.698961938**	0.301038062	1
Pulmonary embolism	0.487804878	0.512195122	1
Gastroesophageal reflux	0.33943662	0.66056338	1
Peptic ulcer disease	0.362600536	0.637399464	1
Esophageal dismotility	0.445897741	0.554102259	1
Acute cholecystitis	0.390134529	0.609865471	1
Acute pancreatitis	0.52617801	0.47382199	1
Costochondritis	**0.808679421**	0.191320579	1
Trauma	**0.768318966**	0.231681034	1
Fibromyalgia	0.286809816	**0.713190184**	1
Neoplasms	0.546531303	0.453468697	1
Rheumatoid arthritis	0.360056259	0.639943741	1
Panic disorder	0.235673931	**0.764326069**	1
Generalized anxiety disorder	0.11717496	**0.88282504**	1
Major depressive disorder	0.329246935	0.670753065	1
Somatic symptom disorder	0.11372549	**0.88627451**	1
Deception syndromes	0.199134199	**0.800865801**	1
Herpes zoster	0.496259352	0.503740648	1
Sarcoidosis	0.274509804	**0.725490196**	1
Substance related	0.548140044	0.451859956	1
Systemic lupus erythematosus	0.166276347	**0.833723653**	1
Acute chest syndrome	0.456896552	0.543103448	1
Total	13.84137174	16.15862826	30

Likelihoods ≥ 0.7 are marked in bold. Each horizontal line separates a unique diagnostic category.

Table S5: Posterior distributions per diagnosis for male patients presenting with chest pain

Diagnosis	Non-Informative	Informative (GP)	Informative (ED)
Myocardial Infarction	0.051196949	0.066518099	**0.216790358**
Angina	0.050339003	0.065403405	**0.163002749**
Pericarditis	0.0449869	0.035069782	0.089644371
Myocarditis	0.042115652	0.010943828	0.052451812
Aortic dissection	0.0412841	0.010727748	0.020566471
Pneumonia	0.038458249	0.049967228	0.067055517
Pleuritis	0.038110385	0.019806105	0.018985424
Asthma/COPD	0.046286233	**0.096220492**	0.046116761
Pneumothorax	0.050498025	0.026244006	0.037734849
Pulmonary embolism	0.035242524	0.009157834	0.008778372
Gastroesophageal reflux	0.024523337	0.050979467	0.006108387
Peptic ulcer disease	0.026196864	0.047651118	0.01957571
Esophageal dismotility	0.032214852	0.025113263	0.01604845
Acute cholecystitis	0.028186117	0.014648427	0.028082917
Acute pancreatitis	0.038014875	0.009878234	0.028406765
Costochondritis	0.058424803	**0.1669997**	0.014552722
Trauma	0.055508874	**0.129816733**	0.027652817
Fibromyalgia	0.020721199	0.016153323	0.005161333
Neoplasms	0.039485342	0.030781013	0.019670385
Rheumatoid arthritis	0.026013047	0.013519075	0.006479451
Panic disorder	0.017026776	0.013273315	0.012723325
Generalized anxiety disorder	0.00846556	0.004399582	0.002108641
Major depressive disorder	0.023787161	0.012362274	0.011850033
Somatic symptom disorder	0.008216345	0.004270064	0.002046566
Deception syndromes	0.014386883	0.003738458	0.003583552
Herpes zoster	0.035853336	0.018633108	0.017861031
Sarcoidosis	0.019832558	0.005153526	0.004939986
Substance related	0.039601569	0.030871618	0.029592429
Systemic lupus erythematosus	0.012012996	0.003121599	0.005984506
Acute chest syndrome	0.033009485	0.008577574	0.016444312
Total	1	1	1

Each horizontal line separates a unique diagnostic category. The most significant results are marked in bold.

Table S6: Posterior distributions per category for male patients presenting with chest pain

Diagnostic Category	Non-Informative	Informative (GP)	Informative (ED)
Cardiovascular	**0.230724008**	0.193492018	**0.545470969**
Respiratory	0.198252968	0.199513005	0.177065911
Gastrointestinal	0.128938249	0.151384032	0.088062591
Musculoskeletal	0.223338627	**0.349623712**	0.082134992
Psychiatric	0.072876182	0.04074409	0.030629643
Other	0.145869967	0.065243144	0.076635893
Total	1.0	1.0	1.0

GP: General Practice, ED: Emergency Department. The highest probabilities are marked in bold.