Mobile health checking is a huge market nowadays, approaching a number of $300 billions. Used to help with diagnosis, information about different medication or with measurements of lab results, these applications can be split in two main categories: intended for patients or intended for doctors and other medical staff. Because our project is part of the latter category we will study the market to see what are the capabillities of the applications available currently.

Epocrates is one such application, used by professional medical staff to quickly get information about medicaments and patient measurements, while also having a feature that allows diagnosis for some given symptoms, yet, some of that functionality is really expensive, worse is that there is no restriction when it comes to users, anyone can download and use the application, which may lead to problems. UpToDate is a product that solves this issue. While it has free content for normal users, UpToDate restricts it's more relevant functionalities, only certified practicioners can access them.

Our project implies two main phases. The collection of existing medical records and the diagnosis of new incomplete records. For the first phase, a doctor would complete an in depth form with relevant data of past cases. The fields of the form are based on the existing categories of real medical records, while loose description will be written in a separate field which will undergo further processing. This processing consists of applying different machine learning algorithms in order to separate a large body of text in a few key attributes. This phase can be evaluated by the numbers of records that we have in storage and by the fidelity of the correlations between our data and the real data added by the doctors.

Only after we have a respectable amount of records in our database we should be able to use the second phase functionality effectively. This phase implies that the user fills up another incomplete form with measurements and symptoms. The application will then use a clustering algorithm to categorise the input with some already existing records, thus giving an exact diagnosis. There are a multitude of highly efficient clustering algorithm including:

* K-means algorithm. One of the basic clustering algorithms, used to categorise n entities to k centers of clusters. In our case these centers would all be different diagnostics. It's main drawback is that it works with a constant number of clusters (k).
* ID3. A machine learning algorithm used to generate a decision tree from a dataset.
* Agglomerative Hierarchical Clustering. Is the most common type of hierarchical clustering used to group entities in clusters based on their similarity.