# Merck medical papers challenge project documentation

Maximilian Balthasar Mansky, Richard Schulz

August 7, 2020

## 1 Introduction

The Merck Medical Papers challenge is a public request to develop an algorithm to best predict future citations from a given article. Dataset for this challenge is the open access PubMed bulk package, around 30 GB of medical papers in XML format.

### 1.1 Data format

In the XML format, the data is organised in a tree structure, with meta data and abstract located under `article/front` and the main text under `article/body`. The `article/back` section contains information regarding funding and citations for the body.

## 2 First Steps

Before delving into the depths of NLP, we first check whether auxiliary information can already predict citations to any accuracy. For this, prepare a table from the downloaded files, including the following headers:

| PMC ID | Journal name | Publication date | Number of citations |
|---|---|---|---|
| `numerical id` | `string` | `date` | `integer` |

The first three can be extracted from the files, the last one needs to captured from the website. This table will also save as a master table for connecting citations to PMC ID.