

# Spá: Extracting and Displaying Attributes of Interest for Evidence Based Medicine

Kuiper, J<sup>1</sup>., Marshall, I.J.<sup>2</sup>, Wallace, B.C.<sup>3</sup>, and Swertz, M.A.<sup>1</sup>

<sup>1</sup> University of Groningen

<sup>2</sup> Kings College London

<sup>3</sup> Brown University

**Abstract.** In an ideal world, scientific findings would be stored in structured databases so as to be easily accessible. In practice, however, unstructured PDF documents remain the main vehicle for dissemination of scientific findings. Those interested in gathering and assimilating data must therefore manually peruse published articles and extract from these the elements of interest. Evidence based medicine (EBM) provides a compelling illustration of this challenge: many expensive person-hours are spent each year extracting summary information from articles that describe clinical trials. Supervised machine learning provides a potential means of mitigating this burden by automating extraction. But for automated approaches to be useful to end-users, we need tools that allow domain experts to interact with, and benefit from, model predictions. To this end, we present an open-source tool called Spá<sup>4</sup> that accepts as input an article describing a clinical trial and provides as output an automatically visually annotated rendering of this article. Specifically, we focus on automating *risk of bias* assessments and extracting the clinical trial sample size. This tool is thus of immediate use to those who practice evidence based medicine. However, more generally, Spá provides a framework for visualizing predictions (both at the document and sentence level) for full-text PDFs.

revision: 70c0be8, date: 2014-04-14

## 1 Introduction

Identifying and extracting specific elements of interest from the full-texts of published articles is an important practical step for many tasks. Consider evidence based medicine (EBM) [?], in which the aim is to address a specific clinical question by identifying and synthesizing data from all published relevant articles. Thus when undertaking such exercises, one needs, e.g., to extract from the article describing a clinical trial the number of participants that were enrolled (the sample size). Another component of EBM is assessing the *risk of bias* for a particular study across different domains. For example, one often wants to assess the risk of bias due to improper blinding of participants and personnel. For this task, one wants both to make a summary judgement (e.g., low risk of bias) while simultaneously extracting the sentence(s) supporting this assessment.

Extracting such information from the free text of articles describing clinical trials is a laborious process. Machine learning methods provide the machinery to automate

---

<sup>4</sup> available under GPLv3 at <https://github.com/joelkuiper/spa>

such extractions: these can effectively impose the structure of interest onto PDF's. But if such technologies are to be practically useful, we need tools to visualize the predicted annotations. Here we describe Spá which aspires to realize this aim. Spá is an open-source, web-based tool that incorporates state-of-the-art machine learning predictors to automatically annotate PDF articles describing clinical trials with risk-of-bias predictions and extracted sample sizes. This tool is useful for practitioners of evidence based medicine and other biomedical researchers.

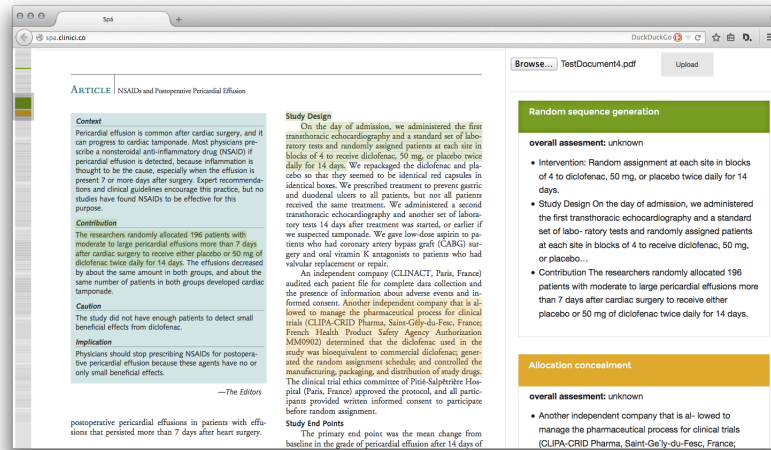


Fig. 1. Screenshot of Spá

While our application of interest is evidence based medicine, we emphasize that the framework can incorporate other types of articles (and the appropriate trained machine learning models). Thus the contribution of this work is two-fold, as we present: (1) a practical tool that incorporates cutting edge machine learning to help biomedical researchers rapidly assess published articles describing clinicals, and, (2) a general open-source web framework for visualizing the predictions of trained models from (PDF's of) full-text articles. These contributions are described further in Sections 2 and ??, respectively.

## 2 Automating Evidence-Based Medicine

### 2.1 Systematic Reviews

The process of pooling and summarizing clinical trials is called *systematic reviewing*, and forms the corner-stone of current evidence based medical practice. Systematic reviewing consists of specifying an inclusion criteria (i.e., the criteria studies must satisfy to be included in the review), searching the literature, screening the retrieved citations to identify eligible studies and, finally, summarizing the relevant evidence.

To aid the process of systematic reviewing we made a generic web-based tool that allows the visualization of annotations within a PDF documents, or meta-information alongside it.

The aim is to have a pluggable system to allow for semi-automated (machine assisted) screening, data extraction and data summarization.

## 2.2 Machine Learning Strategies

- Basically show that we're using state of the art
- Briefly talk about cochrane DB / distant supervision
- KDD stuff (briefly)
- Multi-task stuff!

## 3 Spá Architecture

**Fig. 2.** Sequence diagram of a typical request-response

Spá relies on Mozilla pdf.js<sup>5</sup> for visualization of the document and text extraction. The results of the text extraction are processed server-side by a variety of (pluggable)

---

<sup>5</sup> <http://mozilla.github.io/pdf.js>

pipelines, as outlined in figure 2. Results from these pipelines, which could be complicated machine learning systems, are communicated back to the browser and displayed. For each of the annotations the relevant nodes in the document are highlighted and a custom scrollbar, inspired by substance.io, that acts as a mini-map is projected to show where it resides within the document. The user can then interactively activate and inspect the specific type of results. **todo?**

## 4 Case Study

As a case study we evaluate the automatic assessment of Risk of Bias in clinical trial publications. **todo**

## 5 Conclusion & Future work

We present a web-based tool for interactive visualization of annotations and metadata on PDF documents. This allows users to see the results from machine learning systems within the context of a specific document. Moreover, we present a case study for Evidence Based Medicine by automatically extracting potential Risks of Bias, with supporting sentences.

However, we believe the tool to be useful for a much wider range of text mining and machine learning applications. To increase the generality of the tool work is being done to support a consumer/producer protocol for the pipelines, allowing developers to quickly plug in new systems. Furthermore, work is being done on allowing users to save selected annotations for sharing and off-line, possibly embedded within the document itself.