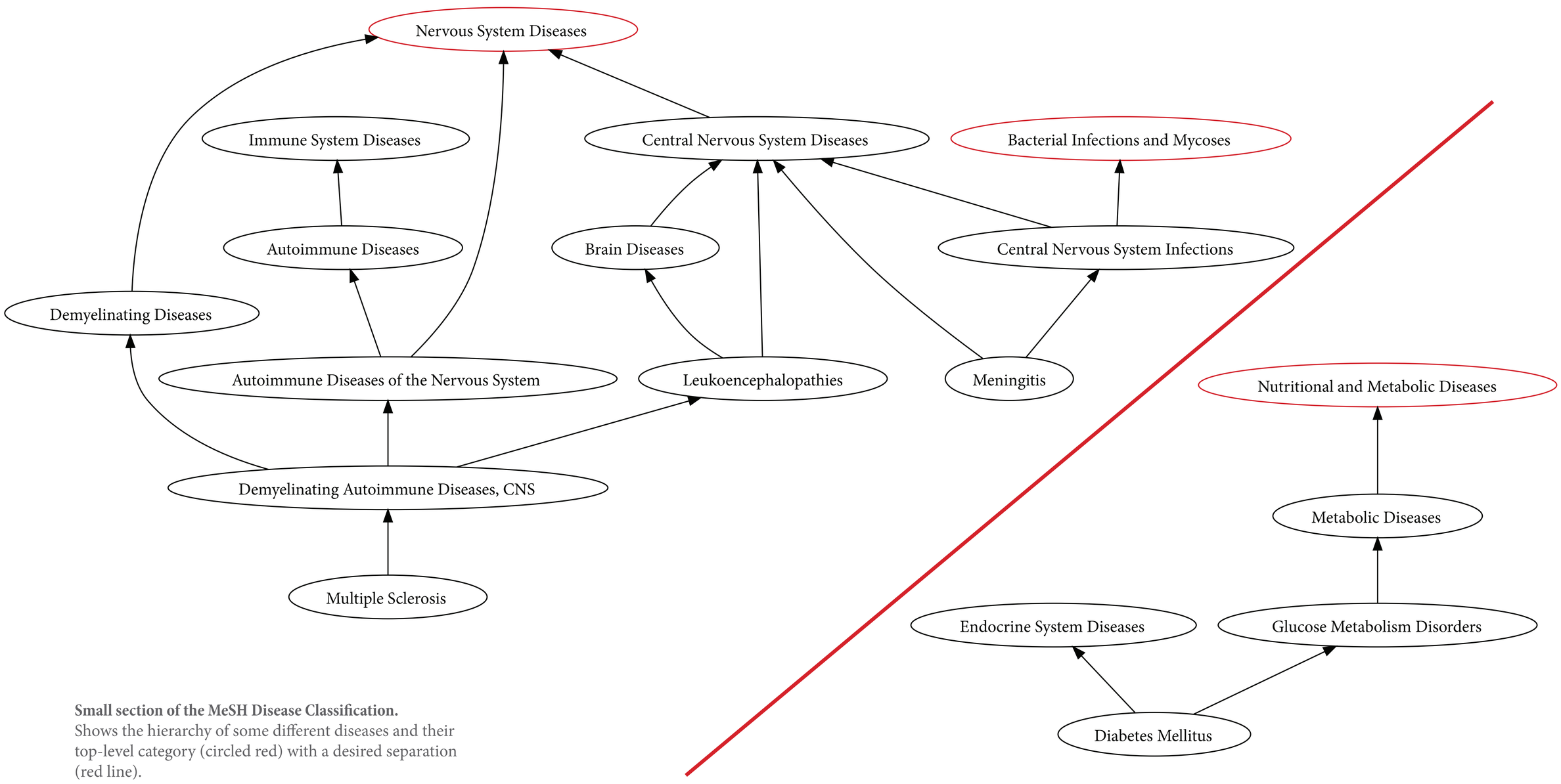
Top-level MeSH Disease Terms Are Not Linearly Separable in Clinical Trial Abstracts

Joël Kuiper & Gert van Valkenhoef
Department of Epidemiology, University Medical Center Groningen, Groningen & Faculty of Economics and Business, University of Groningen

1 Assessments of the efficacy and safety of medical interventions are based on **systematic reviews** of clinical trials. Systematic reviewing requires the screening of vast amounts of publications, which is currently done by hand. In some cases systematic reviewers reduce thousands of publications to only a handful. Whereby, they maximize sensitivity and sacrifice specificity.

2 To reduce the number of publications that need to be screened manually, we tried automated classification of publications by disease category using a **Support Vector Machine** as a supervised classifier.



Small section of the MeSH Disease Classification.
Shows the hierarchy of some different diseases and their top-level category (circled red) with a desired separation (red line).

3 The classification was based on the **ontological structure** of the Medical Subject Headings (MeSH) by treating all terms as their top-level disease category.
For example, Diabetes Mellitus `is-a` Glucose Metabolism Disorder, which is categorized as the MeSH top-level category Nutritional and Metabolic Diseases.

4 The title and abstract of **226,710 publications labeled as "Randomized Clinical Trial"** in PubMed were used as input to the Support Vector Machine. The input text was represented as a bag-of-words and for each of the words its Inverse Document Frequency was calculated.
For all of the 22 top-level categories a one-vs-the-rest classifier was constructed and performance was measured by using a 10-fold cross-validation.

5 Unfortunately, the resulting classifier **lacked sufficient sensitivity** (median of 0.53) for use by systematic reviewers. One explanation could be that for medical terminology the ontological descendants of a top-level item do not sufficiently generalize that top-level item. In that case the Support Vector Machine would fail to find a linear seperating hyperplane.

university of groningen

umcg

Joël Kuiper <me@joelkuiper.eu>
Gert van Valkenhoef <g.h.m.van.valkenhoef@rug.nl>
http://drugis.org/about
22 May 2013