**On the prediction-inference dilemma in biomedicine**

Danilo Bzdok1,2,3,\* Denis Engemann3, Olivier Grisel3, Gaël Varoquaux3, Bertrand Thirion3

1 Department of Psychiatry, Psychotherapy and Psychosomatics, RWTH Aachen University, 52072 Aachen, Germany  
2 JARA-BRAIN, Jülich-Aachen Research Alliance, Germany  
3 Parietal team, INRIA, Neurospin, bat 145, CEA Saclay, 91191 Gif-sur-Yvette, France

\* corresponding author: Prof. Danilo Bzdok, MD, PhD

Email: [danilo.bzdok@rwth-aachen.de](mailto:danilo.bzdok@rwth-aachen.de)

Phone: +49 241 80-85729

Universitätsklinikum Aachen

Pauwelsstr. 30

52074 Aachen

GERMANY

# Abstract

Many achievements of empirical research and evidence-based medicine in the 20th century were grounded in p-values and accompanying methods. In the 21st century, growing ambitions towards precision medicine put a premium on accurate predictions on the single-patient level. This shift of focus incurs tension between established tools to draw statistical inference on the broader population and emerging machine-learning tools to achieve accurate future predictions for particular individuals. Here, we provide an explicit comparison between classical linear regression that identifies significant contributing factors and learning algorithms that automatically select predictive measures. In artificial data simulations and widespread medical datasets, we quantitatively characterized instances when inference and prediction agree and disagree. While both approaches to defining importance in empirical science often allowed for similar conclusions, we describe divergence in a number of data-analysis settings: variables can turn out to be predictive but not significant, or significant but not predictive. More complete understanding of different ways to reach rigorous conclusions from data will be a prerequisite for generating reproducible and clinically exploitable knowledge in biomedicine.

**Keywords**: scientific discovery | statistical significance | prediction performance | variable importance

# Introduction

**Methods**

**Results**

**Discussion**

**Acknowledgements**

DB was funded by the Deutsche Forschungsgemeinschaft (DFG, BZ2/2-1, BZ2/3-1, and BZ2/4-1; International Research Training Group IRTG2150), Amazon AWS Research Grant (2016 and 2017), the German National Merit Foundation, as well as the START-Program of the Faculty of Medicine (126/16) and Exploratory Research Space (OPSF449), RWTH Aachen. The authors declare no competing interests.

**References**

**Figure Legends**

**Figure 1**

**Title. (a)** Abc.