# Dimension reduction

Community health workers in a school, Rabai county, Kenya, September 2021

**WASHA, Takwimu, UKZN, 25 August 2023**

**Till Bärnighausen**, Heidelberg Institute of Global Health, University Hospital and Medical Faculty, Heidelberg University

# Quantitative analysis serve many important functions in health systems research

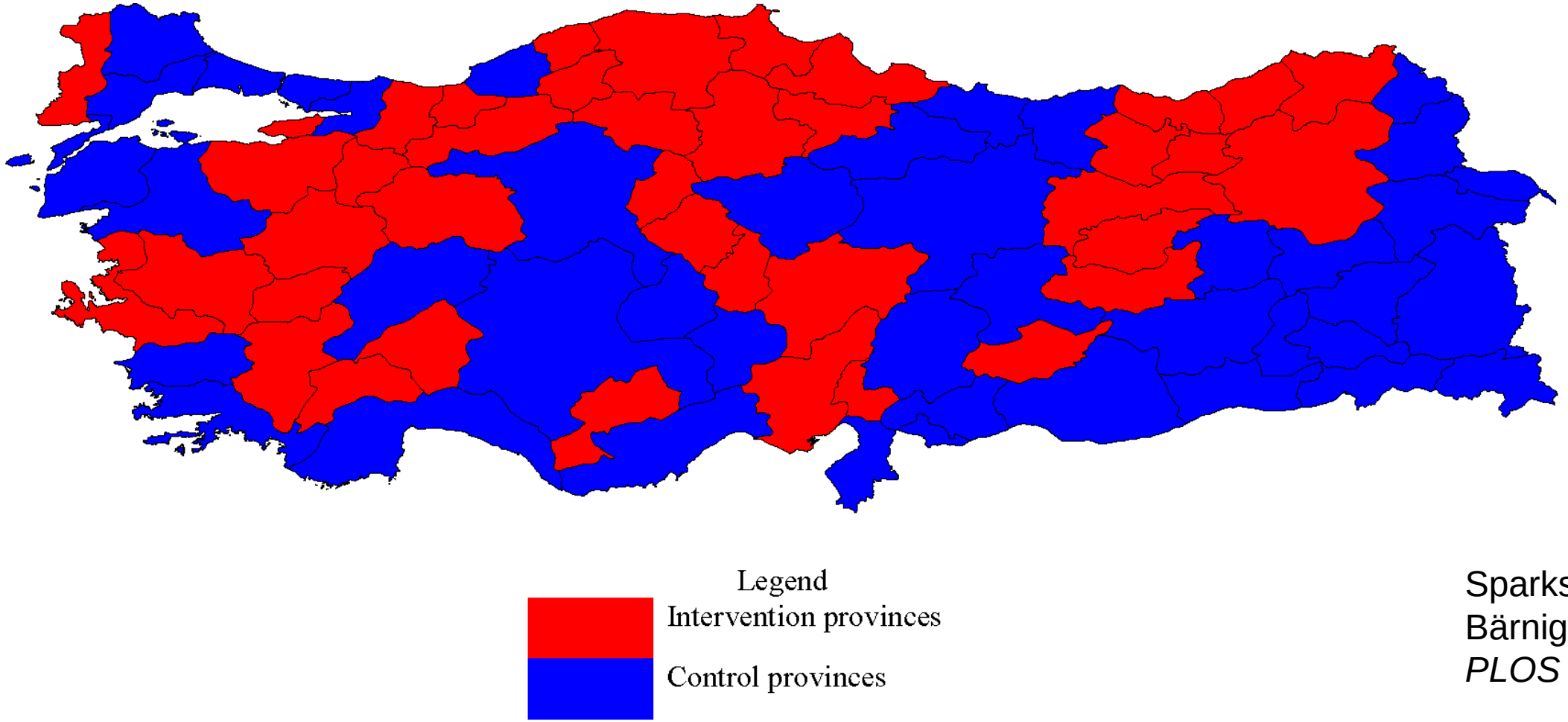
## 4 FUNCTIONS

1. Description
2. **Discovery – unsupervised machine learning**
    - **Dimension reduction**
    - **Cluster analysis**
3. Prediction
4. Causation

HEIDELBERG INSTITUTE OF GLOBAL HEALTH

# Did the Family Medicine Reform in Turkey boost patient satisfaction?

## DIFFERENCE-IN-DIFFERENCES DESIGN, TURKEY




Sparks, Atun & Bärnighausen *PLOS ONE* 2019

# Family medicine reform boosted patient satisfaction

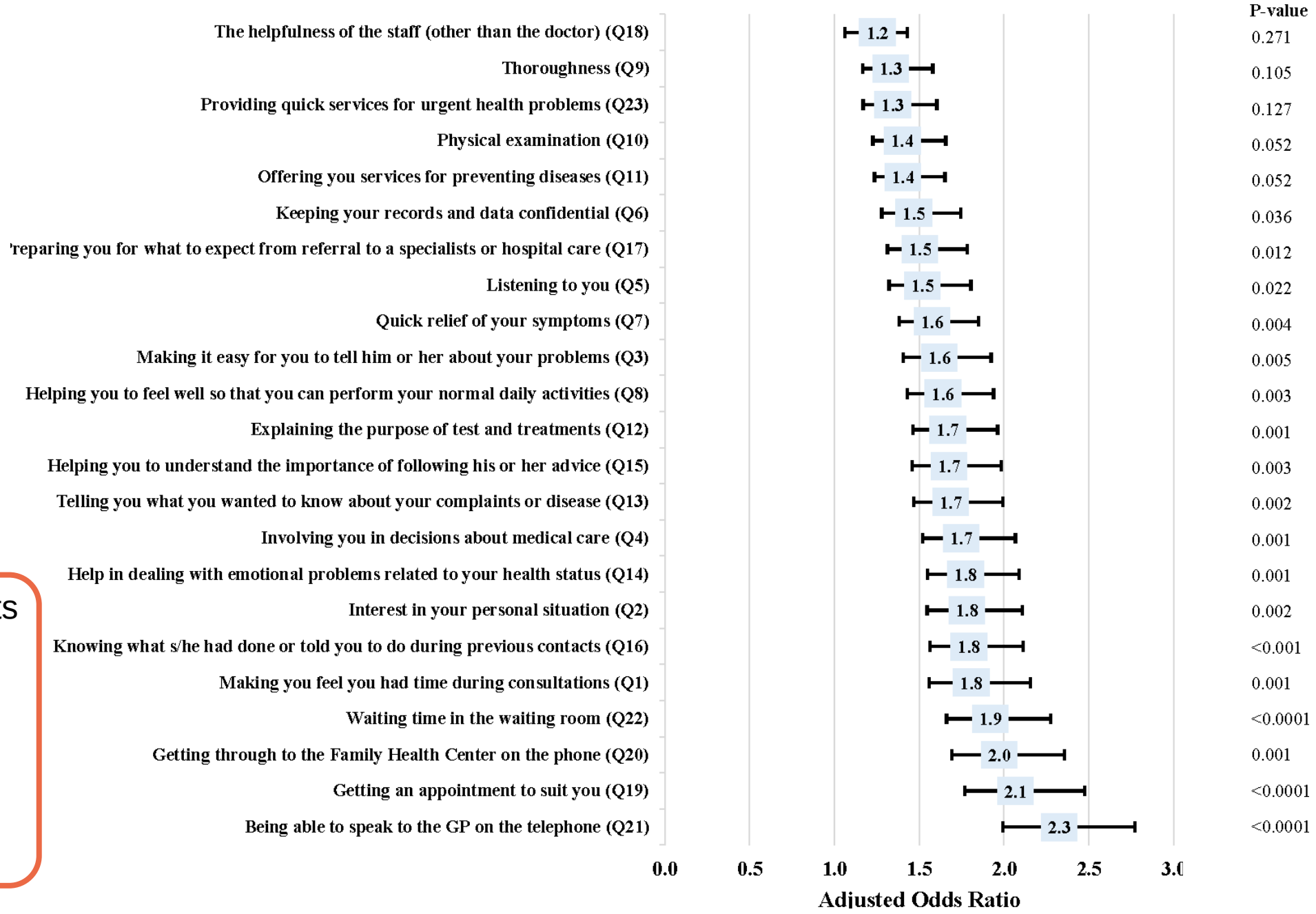
## RESULTS

66,028 primary care patients

2 principal components

**1. Clinical behavior**
**1.99** (95% CI 1.37-2.37)

**2. Organization of care**
**1.18** (95% CI 1.05-1.13)




Sparks, Atun & Bränighausen *PLOS ONE* 2019

# Dimension reduction serves fundamental and technical purposes in data science

## PURPOSES OF DIMENSION REDUCTION

Insight:

- Identifying latent constructs

Data preprocessing:

- Less storage
- Faster computation
- Avoiding algorithmic problems
- Higher accuracy
- More insightful visualization

HEIDELBERG INSTITUTE OF GLOBAL HEALTH

# Dimension reduction serves fundamental and technical purposes in data science

## EXAMPLES

Insight:

- Psychological constructs
- Economic constructs (e.g., wealth index)

Data preprocessing:

- Reducing multicollinearity in regression
- Avoiding the ‘curse of dimensionality’ in clustering analysis

HEIDELBERG INSTITUTE OF GLOBAL HEALTH

# The two-dimensional visual intuition for dimension reduction can be generalized to many dimensions

## PICTURE ON WHITEBOARD

# Principal component analysis is the most common unsupervised dimension reduction approach

## EXAMPLES DIMENSION REDUCTION APPROACHES

Unsupervised:

- PCA, kernel PCA, FAMD
- Low variance filter
- Missing value ratio

Supervised:

- Backward, forward, sequential selection
- Lasso regression
- Random forests

HEIDELBERG INSTITUTE OF GLOBAL HEALTH

# Principal component analysis uses several foundations of statistics

## STATISTICAL INGREDIENTS TO PCA

- Variance and covariance
- Covariance matrix
- Eigenvectors and eigenvalues
- Eigen decomposition
- Singular value decomposition

# PCA is easy to program and easy to implement and artful to interpret

## TECHNICALITIES

- Meaning of eigenvalues
- Explained variance ratio
- Scree plot
- Kaiser’s criterion
- Factor loadings

| Variables | Factor loadings | |
|---|---|---|
| | PC1 | PC2 |
| Income | 0.75 | 0.23 |
| Eduction | -0.83 | -0.15 |
| Risk aversion | 0.12 | 0.89 |
| Future optimism | -0.01 | 0.78 |

HEIDELBERG INSTITUTE OF GLOBAL HEALTH

# Principal component analysis uses several foundations of statistics

## ASSUMPTIONS

1. Linearity
2. Mean and variance sufficient statistics
3. Large variance mean important dynamics (high SNR)
4. Principal components are orthogonal