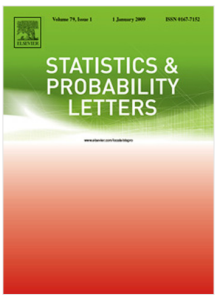

Big data: Some statistical issues

D.R. Cox, Christiana Kartsonaki, Ruth H. Keogh

PII: S0167-7152(18)30060-9
DOI: https://doi.org/10.1016/j.spl.2018.02.015
Reference: STAPRO 8137

To appear in: *Statistics and Probability Letters*

Please cite this article as: Cox D.R., Kartsonaki C., Keogh R.H., Big data: Some statistical issues. *Statistics and Probability Letters* (2018), https://doi.org/10.1016/j.spl.2018.02.015

BIG DATA: SOME STATISTICAL ISSUES

By
D. R. Cox
Nuffield College, Oxford OX1 1NF, UK
Christiana Kartsonaki
Medical Research Council Population Health Research Unit, Nuffield Department of Population Health, University of Oxford, Oxford OX3 7LF, UK
and
Ruth H. Keogh
Department of Medical Statistics, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK

ABSTRACT

A broad review is given of the impact of big data on various aspects of investigation. There is some but not total emphasis on issues in epidemiological research.

1 Introduction

Over the last 125 years computational techniques have evolved from slide rule and log tables, through hand operated machines like the Brunsviga, to electric desk-top machines, and from them to modern computers, at first complex to use and limited in scope and then to the ever expanding modern ubiquitous version. The development of statistical technique and theory over that time has mirrored and been strongly influenced by that growth in computer power and availability.

Big data have been around a long time, for example in population censuses. In an engineering context, paper traces recorded such properties as the stress at various points in an aircraft wing during flight. In a manufacturing context, the mass per unit length of textile yarn was recorded. These examples produced very large amounts of data for visual inspection, but in the past suitable for quantitative analysis at most on a sampling basis. Three questions that characterize today's big data are largely absent from these earlier contexts. In outline the questions are: Are the data relevant for the purpose of the investigation? Is the data quality adequate for its intended purpose? Is the detailed statistical analysis appropriate, in particular