Big Data Epidemiology *Design, conduct and interpret epidemiological studies in large routine healthcare databases*	**Medical Statistics for Big Data** *Advanced statistical methods for the analysis of electronic health records data*
Aims of the module This module aims to provide competence in the design, execution and interpretation of analyses of large databases of routinely collected data, such as the Clinical Practice Research Datalink (CPRD).	**Aims of the module** The module aims to give students the skills needed to analyse data from electronic health record (EHR) studies, including from specialised study designs. Participants will apply these methods using appropriate software. The module will emphasise the practical application of statistical methods to real-life problems, rather than statistical theory. Students will learn what to consider when choosing an appropriate analysis method and how to interpret their results.
Prerequisites Previous statistical experience including the following is a pre-requisite for this module: statistical concepts – p-values, confidence intervals, standard error; multivariate linear, logistic and Cox regression; competence in statistical software including Stata.	**Prerequisites** This module is designed for health care professionals or graduate students who already have some knowledge of designing studies using electronic health record (EHR) data as well as some previous experience of statistical analysis. We expect students to have a similar level of competence as for Big Data Epidemiology (BDE).
Summary of similarities and differences BDE concentrates on the design of EHR studies. Students learn about key issues in EHR studies such as dealing with confounding and missing data. Some statistical competence is required for BDE as students need to be able to understand the concepts and terminology under discussion, but statistical analysis methods are not covered in depth and students do not gain competency in applying them. Medical Statistics for Big Data is complementary in that it aims to give students a greater depth of knowledge of relevant advanced statistical methods as well as experience of applying them to appropriate case studies.	
Learning outcomes • Gain knowledge of examples of large routine healthcare databases in the UK that are available for epidemiological research; • Understand the types of epidemiological questions and study designs where routinely collected data may be of use; • Understand and gain experience, using Stata software, of key practical components involved in designing research in a database, with focus on the CPRD:	**Learning outcomes** • Understand the statistical concepts and assumptions underlying generalised linear models such as Poisson regression. • Be able to apply conditional methods to specialised study designs such as case-crossover designs and self-controlled case-series.

- Identify different populations for analysis and understand which are appropriate to address specific research questions - Identify recorded demographic, clinical history and prescribing data and understand their potential limitations - Gain specific skills in Stata for data extraction and analysis - Gain familiarity of potential data linkages, and their advantages and disadvantages; Gain awareness of data security issues; - Understand relevant methodological issues and possible solutions in routinely collected healthcare data, including: measurement error, missing data, addressing confounding, multiple testing and sample size calculations; - Gain awareness of specialist design and analytical methods relevant to routinely collected data; - Understand the strengths and weaknesses of routinely collected data and use this knowledge to interpret and critically appraise study findings; - Gain the ability to plan a research protocol to answer questions in evidence-based healthcare using routinely collected data.	- Understand the principles of multiple imputation for missing data, when it is appropriate, and how to analyse data after multiple imputation. - Have an awareness of the principles underlying causal inference and how propensity-scores can be used in this context. - Understand how to extend regression approaches to the analysis of interrupted time series. Have an awareness of some of the methodological issues such as autocorrelation. - Be able to choose appropriate statistical analyses for questions using EHR data. - Have gained experience conducting statistical analyses using methods taught on the course and be able to interpret the results.