Analytic infinite-server queueing models are well-established and are capable of predicting many aspects of the time-dependent behaviour of ‘unfettered demand’ for a wide range of realistic situations, including planned and unplanned demands, time-dependent demands and time-dependent service times - in both continuous and discrete time. See for example Massey and Whitt (1993) and Gallivan and Utley (2005).

Such models are directly applicable to real systems which have sufficient servers to avoid queues, and provide valuable approximations for systems which aim to provide high service levels. See for example the use of the Erlang B formula in call centres, and the use of a square root staffing rule to underpin staffing patterns in A&E departments (see Izady and Worthington (2012)).

In this talk we summarise the key queueing models and outline ways in which they can be used to calculate performance indicators (in a hospital setting) which enable comparison between hospitals after adjustment to allow for hospital characteristics reflecting their size, demand level and case-mix.

