Mapping English GP Prescribing Data: a tool for monitoring health-service inequalities

Barry Rowlingson, Euan Lawson, Benjamin Taylor, Peter J. Diggle

Appendix: statistical method for spatial smoothing of prescribing rates

Consider a geographical area \( A \) containing \( m \) GP practices. Let \( x_i \) denote the location of the \( i \)th practice and \( d_i(x) \) the distance between \( x_i \) and an arbitrary location \( x \). Also, let \( N_i \) be the capitation of the \( i \)th practice, \( T_i \) the total cost of prescribed items of interest, and \( Y_i = T_i/N_i \) the observed prescribing rate. Define a smoothing kernel, \( k(d) \), to be a non-negative-valued function of distance, \( d \). For any location \( x \), define \( r(x) \) to be the smallest value such that the sum of the \( N_i \) over all practices located within the disc with centre \( x \) and radius \( r(x) \) is at least \( M \), where the value of \( M \) is to be specified. Then, the smoothed prescribing rate at the location \( x \) is

\[
s(x) = \frac{\{\sum w_i(x)Y_i\}}{\{\sum w_i(x)\}}
\]

where the summation is over all practices and

\[
w_i(x) = k\{d_i(x)/r(x)\}N_i.
\]

For an intuitive interpretation of (1), consider the so-called uniform kernel function,

\[
k(d) = \begin{cases} 1 & : d \leq 1 \\ 0 & : d > 1 \end{cases}
\]

(3)

With this choice of kernel function, equations (1) and (2) define the smoothed prescribing rate as a weighted average of prescribing rates over all practices located within a distance \( r(x) \) of the location \( x \), with a total included capitation of approximately \( M \) and individual practices weighted proportionally to their capitations.

The kernel function (3) therefore leads to a natural interpretation for the smoothed prescribing rates. However, it incorporates an abrupt cut-off of practices included in the averaging, which is intuitively unappealing because catchments for individual practices follow individual patient choices, resulting in a diffuse spatial distribution of patient locations around each practice. For this reason, we prefer to use a quartic kernel function,

\[
k(d) = \begin{cases} (1 - u^2)^2 & : d \leq 1 \\ 0 & : d > 1 \end{cases}
\]

(4)

This has the effect of differentially weighting the contributions of each practice according to their distance from \( x \), with those closest to \( x \) being given the largest weights.

To assess the statistical significance of peaks and troughs in the smoothed prescribing rates \( s(x) \), consider the null hypothesis that the underlying prescribing rates, \( \rho(x) \) say, do not vary spatially, i.e. \( \rho(x) = \rho \) for all locations \( x \). Each \( Y_i \) then has expectation \( \rho \), which we estimate as \( \hat{\rho} = (\sum Y_iN_i)/\sum N_i \), the sample mean of the
observed prescribing rates. Now, make the following two assumptions: firstly, observed prescribing rates vary independently between practices; secondly, the variance of \( Y_i \) is inversely proportional to its capitation, \( N_i \), hence \( \text{Var}(Y_i) = \sigma^2/N_i \). It follows that the sampling variance of the smoothed prescribing rate \( s(x) \) is

\[
v(x) = \sigma^2 \left\{ \sum w_i(x)^2/N_i \right\}/\left\{ \sum w_i(x) \right\}^2
\]

To estimate \( \sigma^2 \), order the \( m \) practices from smallest to largest values of the \( N_i \), group into percentiles and let \( s_k \) and \( \bar{N}_k \) be the sample variance of the \( Y_i \) and the sample mean of the \( N_i \), respectively, within the \( k \)th group. Calculate the least squares regression of \( \log s_k \) against \( \log \bar{N}_k \). Then, \( \log \sigma^2 \) is the intercept of the fitted line. Note also that the scatterplot of \( \log s_k \) against \( \log \bar{N}_k \) also provides a graphical check on the assumption that \( \text{Var}(Y_i) \) is inversely proportional to \( N_i \).

The \( z \)-score to test departure from \( \rho(x) = \rho \) is now

\[
z(x) = s(x) - \hat{\rho}/\sqrt{v(x)},
\]

with \( \hat{\sigma}^2 \) substituted for \( \sigma^2 \) in (5). Under the null hypothesis, each \( z(x) \) is approximately Normally distributed with mean zero and variance one. Hence, for example, the contours \( z(x) = \pm 1.96 \) partition \( A \) into three sub-areas for which the smoothed prescribing rates \( s(x) \) are significantly lower than, significantly higher than, and not significantly different from, the area-wide average. Note that because each \( s(x) \) is estimated from a set of GP practices whose total capitation, \( M \), does not vary spatially, locations with high absolute \( z \)-scores will generally be those with extreme values of \( s(x) \); generally rather than exactly, because the spatial configuration of individual practices around \( x \) also affects the variance of \( s(x) \). Because our method places no prior restriction on the form of the spatial variation in underlying prescribing rates \( \rho(x) \), the assumption of independence between observed prescribing rates \( Y_i \) will usually be reasonable; an exception would be for items prescribed prophylactically to groups of patients perceived to be at risk in a particular region of \( A \). The assumption that the variance of \( Y_i \) is inversely proportional to \( N_i \) is reasonable to the extent that capitations can be taken as proxies for the expected numbers of standard prescriptions of a given item issued per practice.

Using the test statistic (6) it turned out that for most locations, the smoothed prescribing rates \( s(x) \) were significantly different from the national average, \( \hat{\rho} \). To identify locations that differed substantially from the national average, we therefore defined two one-sided test statistics as follows,

\[
z_1(x) = s(x) - c \times \hat{\rho}/\sqrt{v(x)}
\]

\[
z_2(x) = s(x) - (1/c) \times \hat{\rho}/\sqrt{v(x)},
\]

and identifying two sets of locations \( x \) for which \( z_1(x) \) was greater than 1.96 and \( z_2(x) \) was less than \(-1.96 \), respectively; for example, with \( c = 4 \) this identified locations for which the local prescribing rate was more than four times, or less than a quarter of, the national average.