# Assessing the Effect of Air Pollution on Human Health Using Drug Sales Data

Marta Blangiardo* Michela Cameletti† Francesco Finazzi‡

**Abstract**

In the late 1980s and early 1990s, evidence of the effect of air pollution on health started to appear, suggesting association between particulate matter and mortality. Since then, a substantial literature has emerged on health effects from exposure to air pollution, which escalated recently with a study by the International Agency for Research on Cancer which found sufficient evidence to classify air pollution as a leading environmental cause of cancer deaths. In this work we study the association between exposure to outdoor pollution and drug sales for respiratory diseases in England. Exposure is evaluated over space and time at areas where air pollution is measured by monitoring stations and a linear mixed models is used to link exposure with drug sales within the same areas. Risk factors and confounding variables are also considered.

**Key Words:** Population exposure, linear mixed models, England, chronic respiratory diseases.

## 1. Introduction

Understanding the impact of air pollution on population health is a well known problem in epidemiological literature. In the past, concentrations of main airborne pollutants have been linked to hospitalizations (see for instance Lanki et al. (2006), Künzli et al. (2012)), morbidity and mortality in several countries (see for instance Schwartz and Marcus (1990), Dockery et al. (1993)).

In this work, following Sofianopoulou et al. (2013), we propose to assess the impact of air pollution on drug prescriptions for respiratory conditions within the primary health care system in England. It has been shown that bad air quality conditions deteriorate chronic respiratory diseases such as asthma (Brunekreef and Holgate (2002)) and Chronic Obstructive Pulmonary Disease (COPD, Ko and Hui (2012)). Thus when air pollution increases, it should lead to an increase in the prescribing rate of some drugs to treat these medical conditions. Asthma and COPD are mainly treated with drugs based on short acting $\beta 2$-agonists which provide quick relief by opening of the airways. Salbutamol 100mcg, Ventolin 100mcg and Clenil Modulite 100mcg represent about $95\%$ of all prescriptions for $\beta 2$-agonists in England so only these drugs are considered in this work.

## 2. Datasets

Drug prescribing data are released monthly by the English National Health Service (NHS) (http://www.nhs.uk/) for all general practices in England and all drugs. The dataset includes around $11'000$ practices and $30'000$ drugs with about $112$ million drug items prescribed each month. The temporal coverage is from August $2010$ to December $2012$. The following additional time-invariant information are available at the practice level : address, ZIP

*MRC-PHE Centre for Environment and Health, Department of Epidemiology and Biostatistics, St Mary's Campus, Imperial College London, London W2 1PG

†University of Bergamo, Department of Management, Economics and Quantitative Methods, via dei Caniana 2, Bergamo, 24217

‡University of Bergamo, Department of Management, Economics and Quantitative Methods, via dei Caniana 2, Bergamo, 24217

code, number of patiences, patient age distribution, asthma prevalence, pulmonary disease prevalence and multiple deprivation index.

In order to link drug prescribing data with air pollution, the AirBase database maintained by the European Environment Agency (EEA) is considered. AirBase provides pollutant concentration data for multiple pollutants measured by monitoring stations at the European level. In this work, only background-type monitoring stations are taken into account as more representative of the average population exposure. Moreover, daily concentration data are averaged over each month to obtain the same temporal resolution of prescribing data. Although multiple pollutants are available, only nitrogen dioxide ($NO_2$) is used as it is observed at a larger number of spatial locations (with respect to other pollutants) and highly correlated with particulate matter $PM_{10}$. As respiratory conditions deteriorate during seasonal influenza outbreaks, the UK monthly flu index released by the World Health Organization (WHO) is included in the analysis as a confounding variable.

## 3. Linking air pollution with drug prescribing

We propose a spatio-temporal analysis to assess the link between air pollution and rate of drug prescription for respiratory medical conditions. Our hypothesis is that prescription rate increases at times and spatial locations where air quality deteriorates.

In order to assess this association, a key aspect is to evaluate population exposure to air pollution. Monitoring stations are often very sparse in space and physical model outputs of pollutant concentrations may not be readily available at the temporal resolution needed. Statistical models could be used to spatially interpolate pollutant concentrations but local conditions are not easily captured even when covariates are available. In this work, rather than trying to estimate the population exposure at high spatial resolution, we follow a pragmatic approach and we use backgroud-type monitoring stations as representative of the average population exposure within a given area. The respiratory prescription rate of each area is evaluated as the average rate of the general practices within the area. This has the benefit of reducing a possible practice-effect, that is, a given practice (or practitioner) might be more or less prone to prescribe drugs. The next section describes the statistical approach used to link pollutant concentration with prescription rate using spatially aggregated data.

## 4. Statistical model

Let $\mathcal{S} = \{\mathbf{s}_1, ..., \mathbf{s}_N\}$, $\mathbf{s}_i \in \mathcal{D} \subset \Re^2$ be the spatial locations of monitoring stations within the geographic region $\mathcal{D}$ and $\mathcal{Q} = \{\mathbf{q}_1, ..., \mathbf{q}_M\}$, $\mathbf{q}_j \in \mathcal{D}$ the spatial locations of the general practices. It is assumed that $M \gg N$, that is, the number of practices is much larger than that of monitoring stations. For each $\mathbf{s}_i$, $\mathcal{B}_i$ is the smallest circumference of radius $r_i$ centred on $\mathbf{s}_i$ such that the practices within $\mathcal{B}_i$ have a total of at least $n$ patients. At each discrete time $t = 1, ..., T$, the average prescription rate related to monitoring station $i$ is given by

$$y(\mathcal{B}_i, t) = \frac{1}{P_i} \sum_{\mathbf{q}_j \in \mathcal{B}_i} y(\mathbf{q}_j, t) \tag{1}$$

where $P_i$ is the number of practices within $\mathcal{B}_i$ while $y(\mathbf{q}_j, t)$ is the prescription rate of practice $j$, that is

$$y(\mathbf{q}_j, t) = \frac{\sum_{k=1}^{K} d_k(t)}{n_j} \tag{2}$$

where $K$ is the number of drugs considered, $d_k(t)$ are the prescriptions for drug $k$ at time $t$ while $n_j$ is the number of patients of practice $j$.

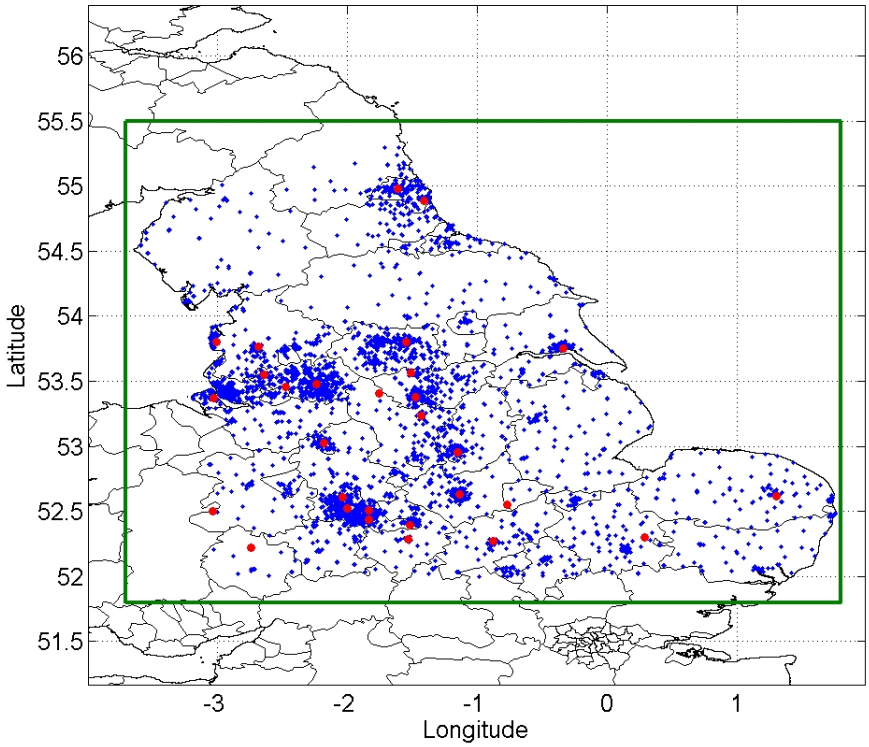


**Figure 1**: Geographic region considered for model estimation. Red dots: $NO_2$ monitoring stations. Blue dots: general practices.

Assuming $y(\mathcal{B}_i, t)$ to be normally distributed, we propose the following alternative linear mixed models:

$$y(\mathcal{B}_i, t) = \beta_0 + x_{NO_2}(\mathbf{s}_i, t)\,\beta_{NO_2} + \mu(\mathcal{B}_i, t) + \varepsilon(\mathcal{B}_i, t) \tag{3}$$

$$y(\mathcal{B}_i, t) = \beta_0(\mathcal{B}_i) + x_{NO_2}(\mathbf{s}_i, t)\,\beta_{NO_2} + \mu(\mathcal{B}_i, t) + \varepsilon(\mathcal{B}_i, t) \tag{4}$$

$$y(\mathcal{B}_i, t) = \beta_0(\mathcal{B}_i) + x_{NO_2}(\mathbf{s}_i, t)\,\beta_{NO_2}(\mathcal{B}_i) + \mu(\mathcal{B}_i, t) + \varepsilon(\mathcal{B}_i, t) \tag{5}$$

where

$$\mu(\mathcal{B}_i, t) = x_a(\mathcal{B}_i)\,\beta_a + x_p(\mathcal{B}_i)\,\beta_p + x_d(\mathcal{B}_i)\,\beta_d + x_f(t)\,\beta_f. \tag{6}$$

For all models, $\beta_0$ is the intercept, $\beta_{NO_2}$ is the parameter related to the pollutant concentration $x_{NO_2}(\mathbf{s}_i, t)$ measured at monitoring station $i$ while $\beta_a$, $\beta_p$, $\beta_d$ and $\beta_f$ are the parameters related to the asthma prevalence $x_a(\mathcal{B}_i)$, to the pulmonary disease prevalence $x_p(\mathcal{B}_i)$, to the multiple deprivation index $x_d(\mathcal{B}_i)$ and to the monthly flu index $x_f(t)$, respectively. Note that, apart from $x_f(t)$ which is given for the whole UK, the other confounding variables in (6) are obtained as average over the practices within $\mathcal{B}_i$. Finally, $\varepsilon(\mathcal{B}_i, t)$ is an i.i.d. normally distributed random error with variance $\sigma^2_\varepsilon$.

Model (4) differs from model (3) as it allows a random intercept $\beta_0(\mathcal{B}_i)$ while model (5) differs from model (4) since $\beta_{NO_2}(\mathcal{B}_i)$ is also random. In all cases, an association between air pollution and prescription rate exists if $\beta_{NO_2}$ is significantly different from zero. When $\beta_{NO_2}$ is random as in model (5), association may exist only for some areas.

## 5. Model estimation

The data presented in Section 2 are here used to estimate models (3-5). Although prescription data are available for the whole of England, only the region $\mathcal{D}$ within bounding box

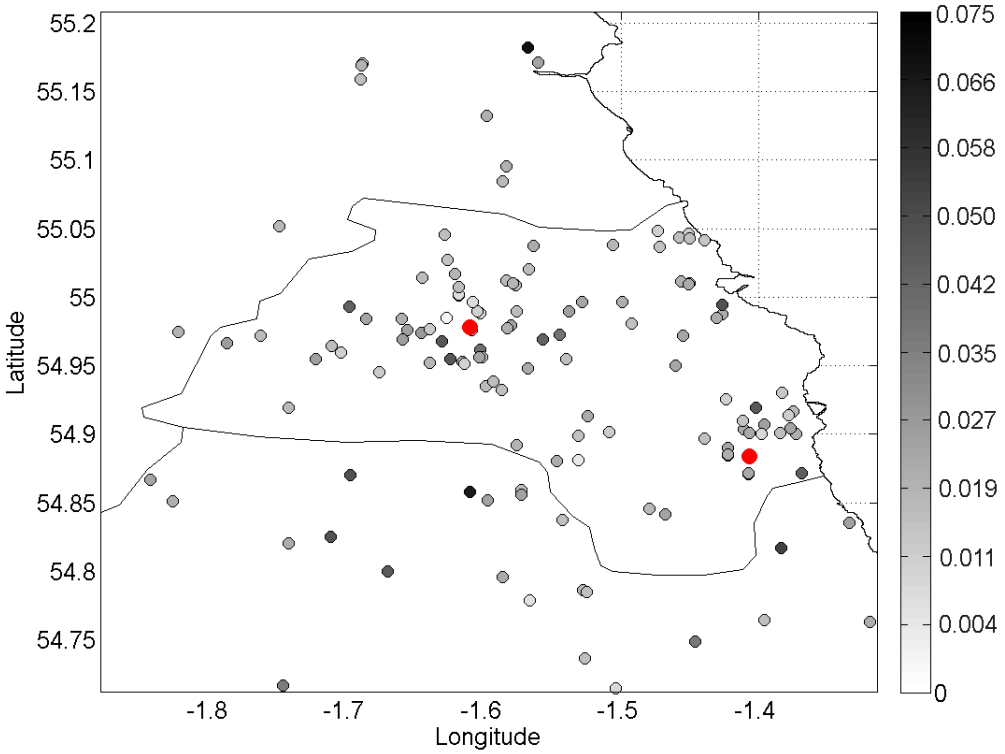


**Figure 2**: Newcastle area. Red dots: monitoring stations. Gray dots: general practice prescription rates $y(\mathbf{q}_j, t)$.

$(55.3°N, 52°S, 6°W, 2°E)$ is considered. This is done to exclude the Greater London area that we believe is characterised by a different dynamic with respect to the rest of England. The above region includes 29 monitoring stations and is depicted in Figure 1, where red dots represent monitoring stations and blue dots are general practices. Note that practices are denser around monitoring stations as the latter are mainly located within cities and urban areas.

Figure 2 shows the prescription rate $y(\mathbf{q}_j, t)$ for the general practices of the Newcastle area. Prescription rate can be quite variable even for practices close in space. Although this behaviour is not fully understood, averaging within circumferences $\mathcal{B}_i$ should mitigate possible problems related to this variability.

Considering $n = 100'000$ patients, the radius of circumferences $\mathcal{B}_i$ ranges from 1.5 to 29.6 $km$ depending on monitoring stations and some of them are depicted in Figure 3. Largest radius are related to more rural areas where the number of practices is smaller.

Model estimation results are reported in Table 1 in terms of log-likelihood and Bayesian Information Criterion (BIC), where lower values correspond to a better fit of the model to the data. The lowest BIC value is related to Model (4) which is characterized by a random intercept and a fixed effect for $NO_2$ concentration. This implies that each area (circumference) has a different average prescription rate but the effect of $NO_2$ concentration is the same across the areas.

Table 2 reports the estimated parameters of Model (4) and their $95\%$ confidence intervals. $NO_2$ concentration, pulmonary disease prevalence and flu index have a positive and significant coefficient while asthma prevalence and deprivation are not significant. Multiple deprivation index is averaged within each circumference and for this reason was expected to be non significant. Asthma prevalence and pulmonary disease prevalence are both re-

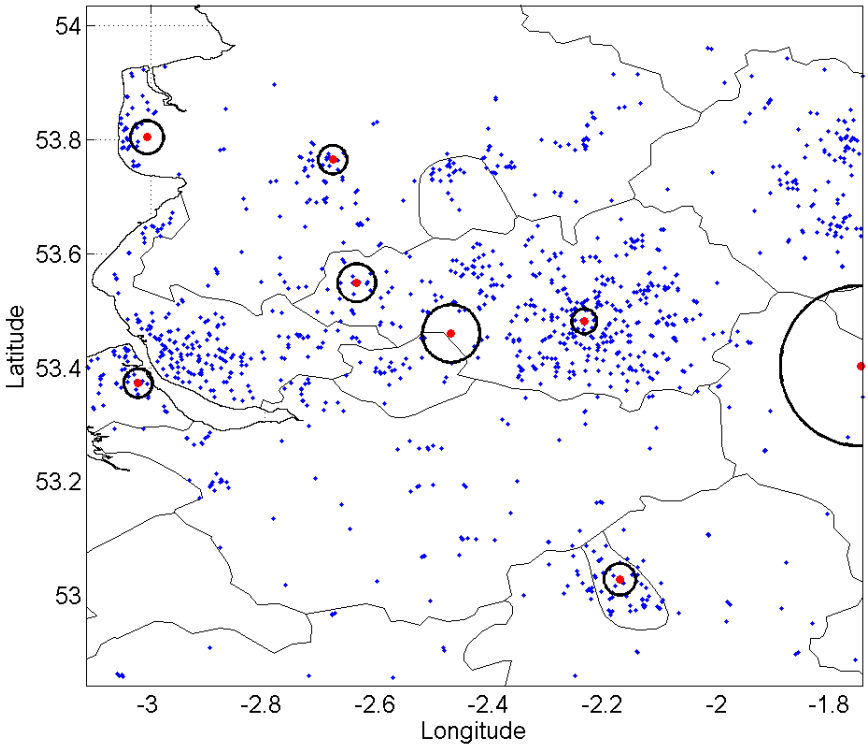


**Figure 3**: Circumferences centred at monitoring station locations.

lated to patients with chronic respiratory diseases and at least one of them was expected to be significant. The estimated $\beta_{NO_2}$ coefficient implies that, when the monthly average $NO_2$ concentration increases by 10 $\mu g/m^3$, prescription rate increases by $1.40\%$ ($95\%$ CI: $0.46\% - 2.25\%$) with respect to an average prescribing rate of $0.03$. This is equivalent to an increase of around $22'000$ prescriptions per month within the geographic region $\mathcal{D}$.

As already discussed, the random intercept implies a different average rate of drug prescription for practices within each circumference. This is shown in Figure 4 where $y^*(\mathcal{B}_i, t) = \hat{\beta}_0(\mathcal{B}_i) + x_{NO_2}(\mathbf{s}_i, t)\,\hat{\beta}_{NO_2}$ is depicted for each circumference $\mathcal{B}_i$. Note that $y^*(\mathcal{B}_i, t)$ is not the actual prescription rate as it does not consider the confounding variables in $\mu(\mathcal{B}_i, t)$.

## 6. Conclusions

The $1.40\%$ variation in rate of drug prescription for each 10 $\mu g/m^3$ increase in the $NO_2$ monthly average concentration is in line with Sofianopoulou et al. (2013). Nonetheless, we believe that the actual impact of $NO_2$ concentration on prescription rate is actually higher and that its effect is partially hidden when monthly averages are considered. Indeed, at least for England, air pollution usually peaks for short periods of time. Misclassification is also

| **Model** | **log-likelihood** | **BIC** |
|---|---|---|
| Eq. (3) | 3179.7 | -6319.1 |
| Eq. (4) | **3731.3** | **-7409.6** |
| Eq. (5) | 3735.0 | -7402.6 |

**Table 1**: Model estimation results.

**Table 2**: Estimated parameters, p-value and and 95% confidence interval for model (4).

| Parameter | Estimate | p-value | Lower | Upper |
|---|---|---|---|---|
| $\hat{\beta}_{NO_2}$ | $\mathbf{4.375 \cdot 10^{-5}}$ | $\mathbf{1.900 \cdot 10^{-3}}$ | $\mathbf{1.613 \cdot 10^{-5}}$ | $\mathbf{7.137 \cdot 10^{-5}}$ |
| $\hat{\beta}_a$ | $1.170 \cdot 10^{-1}$ | $4.494 \cdot 10^{-1}$ | $-1.860 \cdot 10^{-1}$ | $4.210 \cdot 10^{-1}$ |
| $\hat{\beta}_p$ | $\mathbf{8.260 \cdot 10^{-1}}$ | $\mathbf{3.047 \cdot 10^{-5}}$ | $\mathbf{4.390 \cdot 10^{-1}}$ | $\mathbf{1.213} \cdot 10^{0}$ |
| $\hat{\beta}_d$ | $9.493 \cdot 10^{-5}$ | $3.731 \cdot 10^{-1}$ | $-1.100 \cdot 10^{-4}$ | $3.000 \cdot 10^{-4}$ |
| $\hat{\beta}_f$ | $\mathbf{1.670 \cdot 10^{-3}}$ | $\mathbf{3.000 \cdot 10^{-4}}$ | $\mathbf{7.700 \cdot 10^{-4}}$ | $\mathbf{2.500 \cdot 10^{-3}}$ |

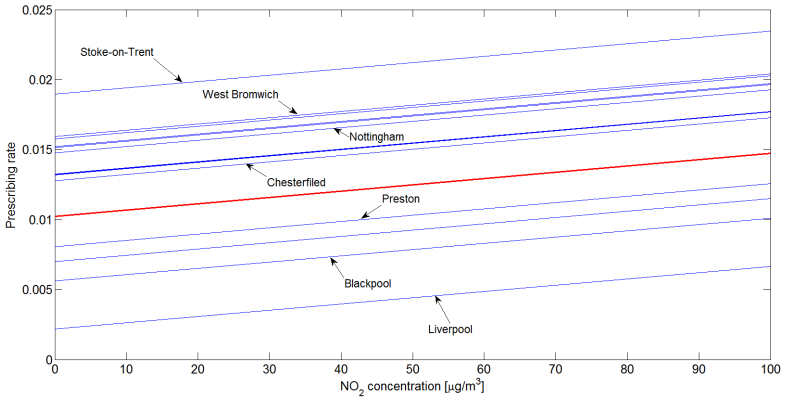


**Figure 4**: Prescribing rate $y^* (\mathcal{B}_i, t) = \hat{\beta}_0 (\mathcal{B}_i) + x_{NO_2} (\mathbf{s}_i, t) \hat{\beta}_{NO_2}$ net of confounding variables for each circumference $\mathcal{B}_i$ (blue lines) and average (red line).

possible, as people do not necessarily live and work (and thus are exposed to air pollution) near the general practice at which they are enrolled. In addition, $NO_2$ concentration and flu outbreaks tend to both peak in winter months leading to potential additional confounding. Future works will try to better model the priscribing rate variability across practices using additional information.

## REFERENCES

Brunekreef, B., and Holgate, S.T. (2002), "Air pollution and health", *The Lancet*, 360, 1233–1242.

Dockery, D.W., Pope, C.A., Xu, X., Spengler, J.D., Ware, J.H., Fay, M.E., Ferris, B.G., and Speizer, F.E., (1993), "An Association between Air Pollution and Mortality in Six U.S. Cities ", *New England Journal of Medicine*, 329, 1753–1759.

Ko, F.W.S., and Hui, D.S.C. (2012), "Air pollution and chronic obstructive pulmonary disease", *Respirology*, 17, 1440–1843.

Künzli, N. (2012), "Is air pollution of the 20th century a cause of current asthma hospitalisations?", *Thorax*, 67, 2–3.

Lanki, T., Pekkanen, J., Aalto, P., Elosua, R., Berglind, N., D'Ippoliti, D., Kulmala, M., Nyberg, F., Peters, A., Picciotto, S., Salomaa, V., Sunyer, J., Tiittanen, P., von Klot, S., and Forastiere, F. (2006), "Associations of traffic related air pollutants with hospitalisation for first acute myocardial infarction: the HEAPSS study", *Occupational and environmental medicine*, 63, 844–851.

Schwarz, J., and Marcus, A. (1990), "Mortality and Air Pollution in London: A Time Series Analysis". *American Journal of Epidemiology*, 105, 1273–1281.

Sofianopoulou, E., Rushton, S.P., Diggle, P.J., and Pless-Mulloli, T. (2013), "Association between respiratory prescribing, air pollution and deprivation" , *Journal of Public Health*, 35, 502–509.