

Chronic Obstructive Pulmonary Disease and Asthma

**Interactive dashboards for different audiences**

Student ID: **10215960**

**IIDS67622** – Understanding data and Decision making

*June 2018*

*“The greatest value of a picture is when it forces us to notice*

*what we never expected to see.”*

*— John W. Tukey*

**Important Note**: The two dashboards are available at: <https://public.tableau.com/views/Maxine_Air_Quality_dashboardv2/Story1?:embed=y&:display_count=yes> (patient)

<https://public.tableau.com/shared/JKQJT3BG5?:display_count=yes> (executive)

Executive summary

Chronic Obstructive Pulmonary Disease (COPD) and asthma are inflammatory chronic conditions that affect the respiratory system. It is estimated that 1 million people suffer from COPD worldwide and 300 million from asthma with the numbers showing growth trends for the next few years. COPD and asthma cost £800 million and £3.7 billion correspondingly in the NHS, putting a significant strain on the healthcare sector.

This report accompanies two interactive dashboards for COPD and asthma for two distinct stakeholder groups; citizens and executives. It has been structured to emphasize how the development stages of visualization tools are inextricably linked to the audiences’ expectations, drawing evidence from the literature and personal experience.

Two different approaches of dashboards are commented:

* the first focuses on citizens’ requirements and highlights the impact of air quality and smoking on COPD and asthma hospital readmissions;
* the second focuses on executive’s requirements, underlining prescription costs of asthma and COPD around England and showing that the North is the most problematic region with prescription costs for 2012 over £197 million.

The dashboards have been developed using free of charge, population data from the NHS Digital and other governmental sources.

The main axes on which this report has been structured are a critical appraisal of the existed health data landscape in the UK and on the potential impact of agile methodology in data science projects; data pre-processing techniques; deterministic data linkage; and data visualization principles on quality and usability of such tools.

Identification of strengths, risks and recommendations complete the narrative that is supported by real-world examples and illustrative graphs.

Table of Contents

[1. Background and review of COPD & Asthma 2](#_Toc516138811)

[2. Health data landscape – Challenges 3](#_Toc516138812)

[3. Decision-making groups 5](#_Toc516138813)

[4. Data visualization tool design – Principles & challenges 6](#_Toc516138814)

[4.1. Principles overview 6](#_Toc516138815)

[4.2 Development process 7](#_Toc516138816)

[4.3 Data acquisition 8](#_Toc516138817)

[4.4 Parsing, Cleaning and Linkage 8](#_Toc516138818)

[4.5 Interpretation 9](#_Toc516138819)

[5. Conclusions – Risks – Recommendations 10](#_Toc516138820)

[References 11](#_Toc516138821)

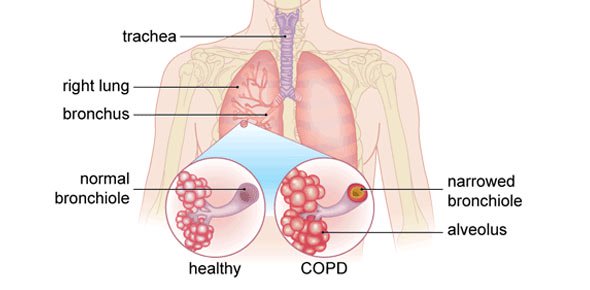
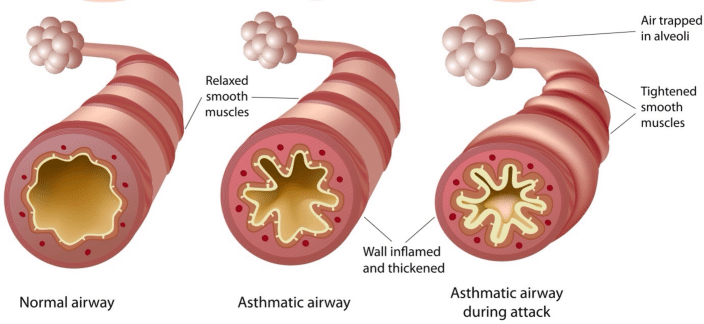
[Appendix 14](#_Toc516138822)

# 1. Background and review of COPD & Asthma

COPD and Asthma are the most common chronic inflammatory respiratory diseases with similar symptoms. The diagnosis is not always easy and can lead a patient to inappropriate medication and uncontrol asthma or COPD, while many people in the UK are undiagnosed (Murphy 2011).

COPD is an “umbrella” term that describes chronic conditions related to the lungs where the patient suffers from respiratory airflow disturbances that are not entirely reversible (WHO 2017). Halbert et al. (2006) in a systematic review and meta-analysis compared 62 studies from 28 different countries and showed that COPD is more prevalent in people with age ≥ 40, in smokers or ex-smokers and in men than in women.

Among the symptoms that characterise COPD we can identify dyspnoea, long-lasting cough, excessive sputum and phlegm production due to inflammation (Figure 1) (WHO 2017; Sharifabad 2017). Smoking, genetic substrate, domestic and environmental pollution are the most common risk factors, while the disease usually affects elderly people (WHO 2017; British Lung Foundation 2015). Asthma represents periods of exacerbation that are characterised by dyspnea and reversible bronchospasm.

**Figure 1**: COPD and inflamed alveolus (left) (Bupa UK 2018) ; Asthma narrowed airways (right) (Quesada and De la Plaza 2017)

According to WHO (2017) COPD is the fifth cause of death worldwide and tends to be the third by 2020 while over 1 million of people in England are living with COPD and 64 million worldwide. Furthermore, deaths in England due to COPD are around 25,000 every year, namely two times the European average (PHE 2015).

Although the disease affects elderly people, it seems to have a great personal, financial and societal impact in working population, turn it into a significant burden. COPD globally corresponds to loss of 27,700 years of productivity per year, placing it from the eleventh to the seventh cause of disease burden (Fletcher et al. 2011). Furthermore, 2% of hospital admissions and more than 3% of bed-days in adults are attributed to COPD, that cost £800 million on the NHS (Rothnie et al. 2016).

Approximately 300 million people globally suffer from asthma that is 4.3% of the population (2014 data) and according to the Global Initiative for Asthma, the above number will be around 400 million by 2025 (Loftus and Wise 2016). Today in the UK there are 5.4 million asthmatics (Asthma-UK 2017) and it is well-known that asthma severity represents fluctuations throughout lifetime, with the highest incidence to be observed in childhood (Nunes et al. 2017).

Sá-Sousa et al. (2014) suggest that it is difficult to measure the prevalence of asthma, due to many ways in which can be defined, thus there is the need for standardised definition. The Asthma UK online survey was conducted in 2017 in a sample of 7,611 people and shows among all that 29.3% of asthma patients do receive unscheduled hospital care in an asthma episode, while 31.5% of them has uncontrolled asthma and 16.6% controlled asthma (Figures 2, 3 - Appendix).

The UK is among the countries with the highest proportion of costs being spend to asthma, with 4 million consultations annually and an average of 330,000 people in primary care being available to treat 25,000 asthmatics. Asthma costs around $5 billion in the UK and 20 million of working-days are lost annually [European Lung white book (2003) citied in Nunes et al. (2017)].

# 2. Health data landscape – Challenges

There is a plethora of available datasets from different sources in the UK that delineate the health data landscape (Figure 7 – Appendix). Sources that support health data science include academia (e.g. UK Biobank); the NHS (e.g. Quality and Outcomes Framework - QOF); the government (e.g. Clinical Practice Research Datalink); public data (e.g. surveys) and other sources such as industries and charities. We could roughly summarise the most important health data sources as follows:

1. Electronic Health Records (EHRs) provide patient level data (e.g. CPRD, The Health Improvement Network – THIN, Qresearch, ResearchOne);
2. Quality of primary care services (QOF);
3. Secondary care sources (Hospital Episode Statistics – HES);
4. Disease registries (e.g. Cancer, Diabetes, Mental Health);
5. Pharmacy and prescribing data for different regions;
6. Census data;
7. Data.gov.uk
8. Deprivation data
9. Mortality (Office of National Statistics) and births datasets;
10. Health surveys (General Practice Patient Survey, National Health Surveys, Annual Asthma Survey);
11. Cohort studies (UK Biobank for genetic data, MRC National Survey for Health and Development);
12. General practice data (General Medical Services, GP Patient Satisfaction);
13. Administrative datasets;
14. Clinical trials
15. Lookup tables
16. User-generated data (e.g PatientsLikeMe, Twitter)

Although the use of data in research has revolutionised epidemiology, it is a field of data science with strengths and weaknesses. Before the analysis process, significant focus should be given on **emerging challenges** related to: data quality; accessibility; complexity; linkage; results generalisability; limitations with geospatial data and population coverage; coding; granularity; privacy and security; data volume; need for computational power; and cost.

The most important characteristics for datasets, identifiers as well as metadata in terms of data **quality** are completeness, correctness, concordance, plausibility and currency (Figure 6, Appendix) (Weiskopf and Weng 2013) and today the NHS aims to standardize the data validation process (NHS England 2016). For instance, primary care data from EHRs suffers from incompleteness, bias due to clinical processes (patients with severe illnesses are more likely to visit a GP or go to hospital), heterogeneity due to different GP management systems in use (EMIS, Vision, SystmOne) and errors, a fact that makes it difficult to have reliable results.

According to Juran (1998) *“high-quality data are data that are fit for use in their intended operational, decision-making, planning, and strategic roles”;* data quality is of great importance as itleads to effective decision-making, lowers financial risks and improves patients satisfaction.

Throughout the visualization tool development process, the data sources that have been used was the QOF, prescription and air quality data from the UK government site, COPD and asthma indicators from PHE[[1]](#footnote-1), population data from the NHS Digital and physical locations of GP practices from the Ordnance Survey. Such data sources are **free of charge**, openly available and there is no great need of privacy and security concerns due to lack of patient level data. In contrast with primary care data sources (CPRD, THIN), such datasets are easily **accessible**, without any specific need for research protocols or payment to acquire access.

In general, NHS Digital **links** administrative data from primary care, from NHS secondary care providers as well as from the social care sector and for some of them retains the right to forward to PHE or NHS Business Service Agency (McNeil 2017). Since 2015, NHS uses the Data Quality Maturity Index (**DQMI**) as data quality indicator to check if the collected datasets meet any standards according to the UK legislation, publishing annually reports (NHS Digital 2018).

QOF[[2]](#footnote-2) was established in 2004, consists of GP practice data and the main purpose is to record chronic diseases quality of care, included asthma and COPD clinical achievement scores per GP practice. It is the largest governmental financial initiative ever took place in primary clinical care (Roland 2004). Interestingly, QOF gives the ability to create a unique dataset that allows the **stratification** of data per GP practice, Clinical Commissioning Groups (CCGs), regions and area teams, providing to users the opportunity to retrieve information in different levels of granularity.

The above datasets cover different **time periods** and sometimes different **geographic locations**, a fact that affects potential linkage with more than one datasets. Moreover, these datasets differ in **granularity**, e.g. EHRs refer to patient level while QOF or prescription data on GP practice level. Although EHRs provide detailed insight on patient level data, covering a wide time period (e.g. CPRD and THIN since 1987 across the UK, QResearch for UK since 1990), for this visualisation tool we have focused on population level data and open-to-public data sources have been used, even though this is a limitation (see part 5).

However, when data **linkage** is essential, differences in granularity, time-periods, geographic locations create records with **missing data**. Problems arise when the reason of missingness is not completely known and the researcher needs to develop the visualisation tool based on assumptions about the missing pattern. Furthermore, it is not always possible to be aware of data cleaning procedures that have already taken place with specific datasets we have used.

# 3. Decision-making groups

This data visualisation tool has been developed to satisfy different needs and requirements of two very distinct decision-making groups; public and professional stakeholder groups.

For data scientists, it is often difficult to take a step back and imagine how things may look if we do not know anything about them, i.e. it is difficult to imagine what someone expects to see in a dashboard, when he/she do not know what a dashboard actually is! This was exactly what we tried to figure out for the public stakeholder group; how citizens imagine a visualization tool, what are their expectations and what they would like to see and explore with such a tool.

The most important message we get from this meeting was to keep things simple and comprehensive, by avoiding scientific terms and complicated concepts and to focus on giving a message for their health condition. Interestingly, it was highlighted that they would like to know the impact of common factors in their life.

The professional stakeholder was interested in asthma and COPD prescription costs around England. He highlighted the need of making comparisons between different regions and the identification of drivers behind the results. The main challenge here was that a great amount of information needs to be packed in order to give one clear message. This was achieved by adding in each page many layers of different functionality, providing the user selection tools and the ability to ask many questions in the same dashboard.

Studies from the ONS suggest that the prevalence of asthma and COPD around England depends on population age, health life expectancy, socioeconomic factors, alcohol consumption, smoking and dietary facts. Furthermore, in 2017 30% of the population live in the North of England, which is home of 50% of the poorest areas (ONS 2017). The prevalence has been included in the dashboards to guide the stakeholder in future decisions, taking into consideration that the cost is affected by the prevalence, and the prevalence from all the above drivers.

Comparing and contrasting the needs of the two stakeholders, we observe that in terms of patients, the main purpose was the data presentation in an interesting, meaningful, simple way. For the manager, a more sophisticated approach was expected; the data should (at least partly) reveal the reasons *why* the cost was high or low in order for the decision-making process. The public groups need and consequently, their decision-making process were related to factors that a potential change could have a positive impact on their quality of life; for patients suffer from either asthma or COPD, the dashboard could be a motivation for change while for the general public a prevention tool. For the executive, the dashboard could reveal patterns to save money by searching for new, cost-effective drugs, taking actions in regions where overprescribing is observed, having a better remote control around the country. Their intentions are the same – *a time for change* – but from a completely different perspective.

# 4. Data visualization tool design – Principles & challenges

## 4.1. Principles overview

According to Few (2007), a dashboard is a visual representation of the most important information, that is arranged on a single screen in the most comprehensive way that helps someone to capture data insights at a glance.

During the design process, data visualisation principles were followed, that are proposed by well-recognised scientists such as Edward **Tufte**, William **Cleveland** and Stephen **Few**. In detail, some of the most important data principles that we have incorporated in the the visualization tool are the following (Tufte 2001; Cleveland 1994; Few n.d.):

* A graph should have appropriate amount of information and should be able to stand alone, avoiding overwhelming visualisations.
* Whenever numbers are used, they should reflect the numerical quantities that have been measured.
* The labelling and legends should be clear without letting questions of uncertainty and ambiguity arise.
* Three-dimensional graphs should be avoided, especially when information could be described in two-dimensions.
* Data variation should be clearly driven and stay in the centre of our focus against design variation.
* The use of monetary measurements is usually preferable than nominal units.
* There is always the need of setting a context and develop the visualisations in this context, omitting any irrelevant graph, especially when this omission does not reflect on information loss.
* Everything that does not add any point in comprehension should be removed – e.g. Tufte (2001) has used the term “*chartjunk*” to describe grids, patterns and graphs that have artistic rather than informative reason of existence.
* Graphics should be indicated by the data nature; horizontal graphs are easier to read, especially when there are long, categorical names (e.g. drug names, Figure 11 – Appendix).
* A major conclusion should be included in some way in the data area.
* The use of scales is important, especially when there is the need of comparisons.
* Diverse view of the same data provides different insights (Figures 9, 10 – Appendix);
* The visualization tool should provoke the user to ask “*why*” and explore the data, focusing on things that are important.

In addition to the above statements, color have been used effectively. For the North of England where the highest amount of money has spent on prescribing, red has been used to highlight that this is a problematic area. In an analogue way the same concept has been applied on the patient dashboard to distinguish smoking prevalence and carbon dioxide ranges, creating a visual hierarchy of the information (Figures 14, 16 – Appendix).

Whenever there is text in the data area, it has been placed on the top left side as it contains instructions and some of the most important inferences. The position has been chosen because this is the way we read, from the top left side with zig-zag motions through a screen (Knaflic 2015). For example, in the executive’s dashboards, the map is the second most important part of the data we want the user to explore so it has been placed exactly after the instructions. Finally, the components in the executive’s dashboard have been placed in three columns and a horizontal space at the bottom. For the patient’s dashboards a simplified approach of two columns has been used.

## 4.2 Development process

Data science projects are usually non-linear and exploratory by their nature, namely susceptible to requirement changes. This fact allows software engineering methodologies to be implemented effectively in a data science pipeline.

**Figure 2:** Common data science project pipeline (Tetali 2017)

Figure 3 shows the classic waterfall methodology and a more agile approach. In a waterfall approach, the role of every team member is fixed while in the agile, the workload can be significantly reduced through frequent iterations and modifications according to updated requirements (Janaitis 2017).

Throughout the development process of this visualization tool, two meetings with the two stakeholders took place until the dashboards final shape. During this process, some of the requirements and priorities changed. We went through this stage, working closely with the stakeholders, making questions and changing roles many times from the planning stage to design and deployment with reflection at regular time intervals. The advantage of this approach is the ability to interexchange opinions and **save resources**, as there is no need to go again through everything whenever there is a new, or slightly changed requirement.



**Figure 3:** Traditional waterfall methodology (left) vs agile methodology (right) (Janaitis 2017)

## 4.3 Data acquisition

All the data have been acquired from the sources mentioned in part 2. In order to develop the patient dashboard, air quality data from the government’s site[[3]](#footnote-3) have been used. The data is **open** for public use and **free** of charge, available for the period 2005-2013 and they contain information about carbon dioxide emissions.

The data about the executive dashboard is based on prescription[[4]](#footnote-4) data that have been acquired from the same source as above for the period 2012-2013[[5]](#footnote-5), as our aim was to link this dataset to the initial[[6]](#footnote-6) one. In order to enable comparisons on prescriptions in different granularity level (CCG, area, region), population data has also been used from the NHS Digital for the same time period, representing the number of enrolled patients[[7]](#footnote-7) per CCG.

## 4.4 Parsing, Cleaning and Linkage

All data manipulations and initial visualizations to test the data quality (e.g. variable types, outliers, inconsistent values – data normalisation, appropriate dosage forms for the drugs, dates to represent only the month throughout the year 2012 etc.) have been done with R.

For the professional dashboard, the cleaning part was easier than for the public dashboard, showing that the data had already been cleaned by the provider. For the practice code and CCG code, a collation was made based on different datasets to reassure that the fields that would have been used as keys to link the datasets were correct and real.

A **deterministic** approach of linkage has been followed for all the datasets, which is a **simple**, **fast**, reliable method for good quality data, minimising the number of false matching but sacrificing quality due to missingness and sensitivity (Zhu et al. 2015).

Before the linkage, the size of the initial dataset was reduced drastically, keeping only the appropriate for the visualization tool variables, without impact on information loss. This affect the tool future **reusability** and **extension**, although it allows us to handle the data easier.

Professional stakeholder dashboard

The linkage for this dashboard followed two steps: a) the linkage between enrolled patients per CCG dataset with the initial one and b) the linkage between the drugs and the dataset from (a). All the initial datasets did not contain missing values. For the linkage, it has been used the dplyr::left\_join() R function. This means that missing values were created however, the aim was to calculate the percentage of missingness before following any strategy for missing data.

The missingness occurred because in the dataset with drugs, there were fewer practice codes than in the initial one. Having checked these practices, the **information loss** in case of removing them from the final dataset would be approximately 13.12%. We decide to continue with complete case analysis. This could potentially impact the **decision-making** process and results **generalisability**, leading to over or underestimation but whenever such assumptions are made due to unknown missing patterns, the results are explained under these assumptions.

The final dataset was an enormous dataset consists of more than 2 million rows. The huge **volume** makes any **manipulation** - and consequently visualisation – difficult and slow, especially when **computational** **power** is limited.

In general, data linkage is one of the reasons that causes **data quality** issues (Christen 2012) that have mentioned in part 2.

Public stakeholder dashboard

The air quality dataset was significantly smaller and thus easier to handle however, the cleaning and linking process were quite difficult. The difficulty lies in the absence of a field that could be used as a key to link air quality data with the initial dataset.

A two-step process was followed: a) the GP practices postcodes from the initial dataset were matched with a lookup of postcodes by local authority and b) this merged dataset was linked to air quality emissions dataset on local authority level. The **matching was** **manually** which indicates the potential of errors that could affect the data quality.

## 4.5 Interpretation

The interpretation of the results is a process that resembles the development pipeline. We need to **revisit** the original question to assess results nature that lead to interpretation. It is a cycle that requires thoroughly thinking on whether the results match the expectations and vice versa, considering any possible implication. Although it may seem a time-consuming process, it is able to save significant amount of resources (energy, money, time, reputation).

During the interpretation, all the followed process of analysis should be considered. For instance, we have presented the results of the dashboards in the general framework that all the assumptions and limitations have settled (e.g. missing data, population level data). The is a clear message for the executive that is the problematic North and an explorative representation for the public stakeholder.

# 5. Conclusions – Risks – Recommendations

The proposed solution is complete and incorporates well all the stakeholders’ requirements. It has been developed based on visualization principles and the main strength is the national scale and the different granularity levels that provide detailed insights into data. The good quality and huge volume of the used datasets reflect how carbon dioxide emissions and smoking affect COPD and asthma in different regions. Furthermore, the cost analysis around treatment provides the executive with a visual tool for having financial control around the country at a glance, supporting decisions in terms of generic drugs or generating new questions around specific regions with highly observed prescription costs.

However, there are limitations that could be improved; for example, the lack of patient-level data induces uncertainties on whether every prescribed item was actually prescribed for COPD and asthma, especially in terms of corticosteroids as they are broad-spectrum drugs. Moreover, relating to the patient’s dashboard, there is no specific pattern that its repetition could lead to a clear causal relationship, especially between carbon dioxide emissions and respiratory diseases; the relationship is quite clearer between smoking and admissions, showing that heavier smokers had more admissions but again there is a proportion of uncertainty due to the use of population-level data. New linkages could provide a data-driven evidence on the results.

Further analysis on how data missingness affects the quality of information could be implemented through comparisons with the same dataset before and after data cleaning or with external datasets. Another point that could have been improved would be the use of probabilistic linkage instead of deterministic approaches, to test whether the quality of the linkage would have been significantly improved, as the literature shows better results for such approaches (Randall et al. 2013).

These dashboards at the moment should be used as hypothesis generating tools for exploring further the reasons why something is happening, generating new discussions around COPD and asthma quality of life improvements and resources saving analysis. A recommendation for further extension would be the incorporation of real-time, streaming data in dashboards that could change at regular time intervals, reflecting the current picture around the country.

# References

Asthma-UK. (2017). Annual asthma care survey. *Asthma UK*. [online]. Available from: https://www.asthma.org.uk/get-involved/campaigns/publications/survey/.

British Lung Foundation. (2015). What is COPD? *British Lung Foundation*. [online]. Available from: https://www.blf.org.uk/support-for-you/copd/what-is-it.

Bupa UK. (2018). Chronic Obstructive Pulmonary Disease. [online]. Available from: https://www.bupa.co.uk/health-information/directory/c/copd.

Christen, P. (2012). *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Berlin Heidelberg: Springer-Verlag. [online]. Available from: //www.springer.com/gb/book/9783642311635.

Cleveland, W.S. (1994). *The elements of graphing data*. Rev. ed. Summit, N.J.: Hobart Press.

Dohyung, K. (2017). Understanding COPD. *Dohyung Kim*. [online]. Available from: http://dohyungkim.com/copd/.

European Lung white book. (2003). The economic burden of lung disease - ERS. [online]. Available from: https://www.erswhitebook.org/chapters/the-economic-burden-of-lung-disease/.

Few, S. (2007). Dashboard Confusion Revisited. *Visual Business Intelligence Newsletter*, p.6. [online]. Available from: http://perceptualedge.com/articles/visual\_business\_intelligence/dboard\_confusion\_revisited.pdf.

Few, S. Data Visualization: 8 Core Principles. *Tableau Software*. [online]. Available from: https://www.tableau.com/blog/stephen-few-data-visualization.

Fletcher, M.J. et al. (2011). COPD uncovered: an international survey on the impact of chronic obstructive pulmonary disease [COPD] on a working age population. *BMC Public Health*, 11, p.612. [online]. Available from: https://doi.org/10.1186/1471-2458-11-612.

GHDx. (2016). GBD Results Tool | GHDx. [online]. Available from: http://ghdx.healthdata.org/gbd-results-tool.

Halbert, R.J. et al. (2006). Global burden of COPD: systematic review and meta-analysis. *The European Respiratory Journal*, 28(3), pp.523–532.

Janaitis, A. (2017). Using Agile Data Science Methods to Manage Shifting Priorities. [online]. Available from: https://www.elderresearch.com/blog/agile-data-science-manages-project-priorities.

Juran, J.M. and Godfrey, A. (1998). *Juran’s Quality Handbook*. 5th ed. McGraw-Hill Companies.

Knaflic, C.N. (2015). *Storytelling with Data: A Data Visualization Guide for Business Professionals*. Wiley. New Jersey.

Loftus, P.A. and Wise, S.K. (2016). Epidemiology of asthma: *Current Opinion in Otolaryngology & Head and Neck Surgery*, 24(3), pp.245–249. [online]. Available from: http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage&an=00020840-201606000-00014.

McDonnell, L., Delaney, B. and Sullivan, F. (2017). *Datasets that may be of interest to Primary Care Researchers in the UK*. The Farr Institute UK. [online]. Available from: http://www.farrinstitute.org/wp-content/uploads/2017/10/Datasets-that-may-be-of-interest-to-Primary-Care-Researchers-in-the-UK-May-2016.pdf.

McNeil, K. (2017). *Review of Public Health England’s data collection and data management functions*. [online]. Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\_data/file/660214/McNeil\_PHE\_data\_collection\_review.pdf.

Murphy, A. (2011). Knowing the differences between COPD and asthma is vital to good practice. *Pharmaceutical Journal*. [online]. Available from: https://www.pharmaceutical-journal.com/learning/learning-article/knowing-the-differences-between-copd-and-asthma-is-vital-to-good-practice/11085597.article.

NACAP. (2016). National Asthma and COPD Audit Programme (NACAP). *Royal College of Physicians*. [online]. Available from: https://www.rcplondon.ac.uk/projects/national-asthma-and-copd-audit-programme-nacap.

NHS Digital. (2018). Current Data Quality Maturity Index (DQMI). *NHS Digital*. [online]. Available from: https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/data-quality.

NHS England. (2016). Data quality improvement. [online]. Available from: https://www.england.nhs.uk/data-services/validate/.

Nunes, C., Pereira, A.M. and Morais-Almeida, M. (2017). Asthma costs and social impact. *Asthma Research and Practice*, 3, p.1. [online]. Available from: https://doi.org/10.1186/s40733-016-0029-3.

ONS. (2017). Overview of the UK population - Office for National Statistics. [online]. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/articles/overviewoftheukpopulation/july2017.

PHE. (2015). Chronic smoking-related lung disease blights over 1 million lives in England. *GOV.UK*. [online]. Available from: https://www.gov.uk/government/organisations/public-health-england.

Quesada, D. and De la Plaza, P. (2017). Time series analysis and forecast of respiratory conditions in Florida. In p. 5103.

Randall, S.M. et al. (2013). The effect of data cleaning on record linkage quality. *BMC Medical Informatics and Decision Making*, 13(1). [online]. Available from: http://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/1472-6947-13-64.

Roland, M. (2004). Linking Physicians’ Pay to the Quality of Care — A Major Experiment in the United Kingdom. *New England Journal of Medicine*. [online]. Available from: https://www.nejm.org/doi/full/10.1056/nejmhpr041294.

Rothnie, K. et al. (2016). *COPD prevalence model for small populations: Technical Document produced for Public Health England*. National Heart and Lung Institute and Department of Primary Care & Public Health, School of Public Health.

Sá-Sousa, A. et al. (2014). Operational definitions of asthma in recent epidemiological studies are inconsistent. *Clinical and Translational Allergy*, 4, p.24. [online]. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4136946/.

Sharifabad, M.A. (2017). COPD - Symptoms, diagnosis and treatment | BMJ Best Practice. [online]. Available from: https://bestpractice.bmj.com/topics/en-gb/7.

Tetali, R. (2017). Agile Data Science. [online]. Available from: http://rao.tetali.info/2017/05/agile-data-science.html.

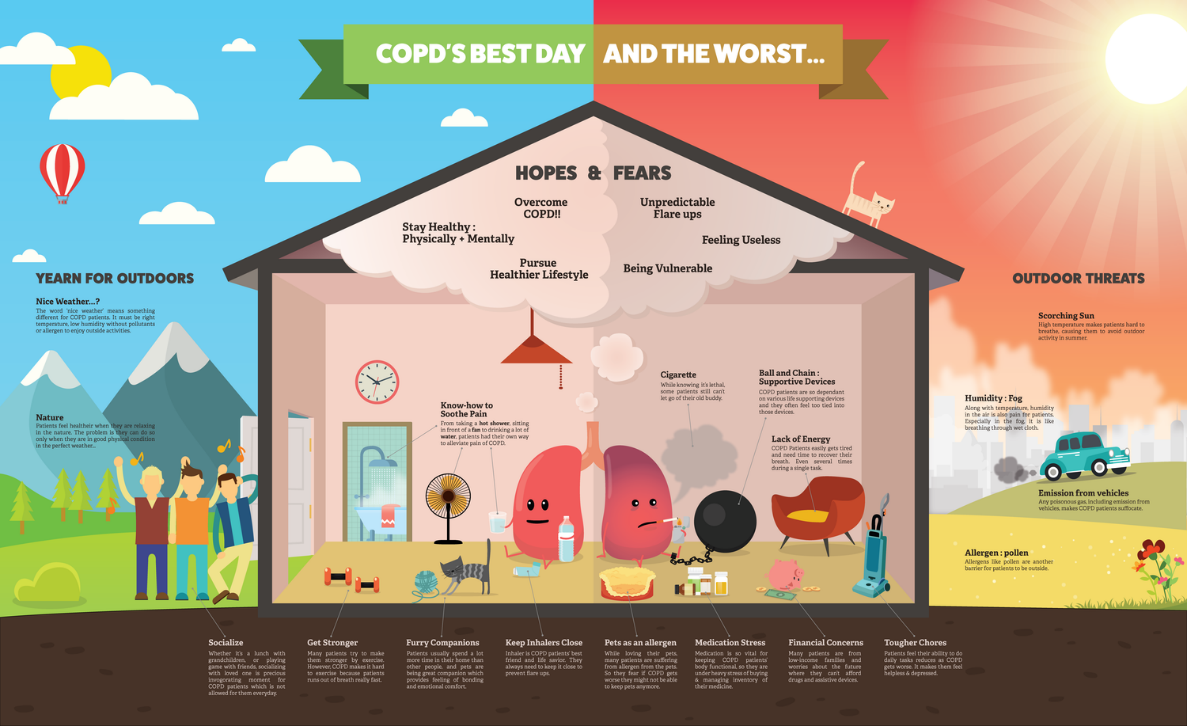
Tufte, E.R. (2001). *The Visual Display of Quantitative Information*. 2nd ed. Cheshire: Graphics Press.

Weiskopf, N.G. and Weng, C. (2013). Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *Journal of the American Medical Informatics Association*, 20(1), pp.144–151. [online]. Available from: https://academic.oup.com/jamia/article-lookup/doi/10.1136/amiajnl-2011-000681.

WHO. (2017). Chronic obstructive pulmonary disease (COPD). *WHO*. [online]. Available from: http://www.who.int/respiratory/copd/en/.

Zhu, Y. et al. (2015). When to conduct probabilistic linkage vs. deterministic linkage? A simulation study. *Journal of Biomedical Informatics*, 56, pp.80–86. [online]. Available from: http://www.sciencedirect.com/science/article/pii/S1532046415000921.

# Appendix



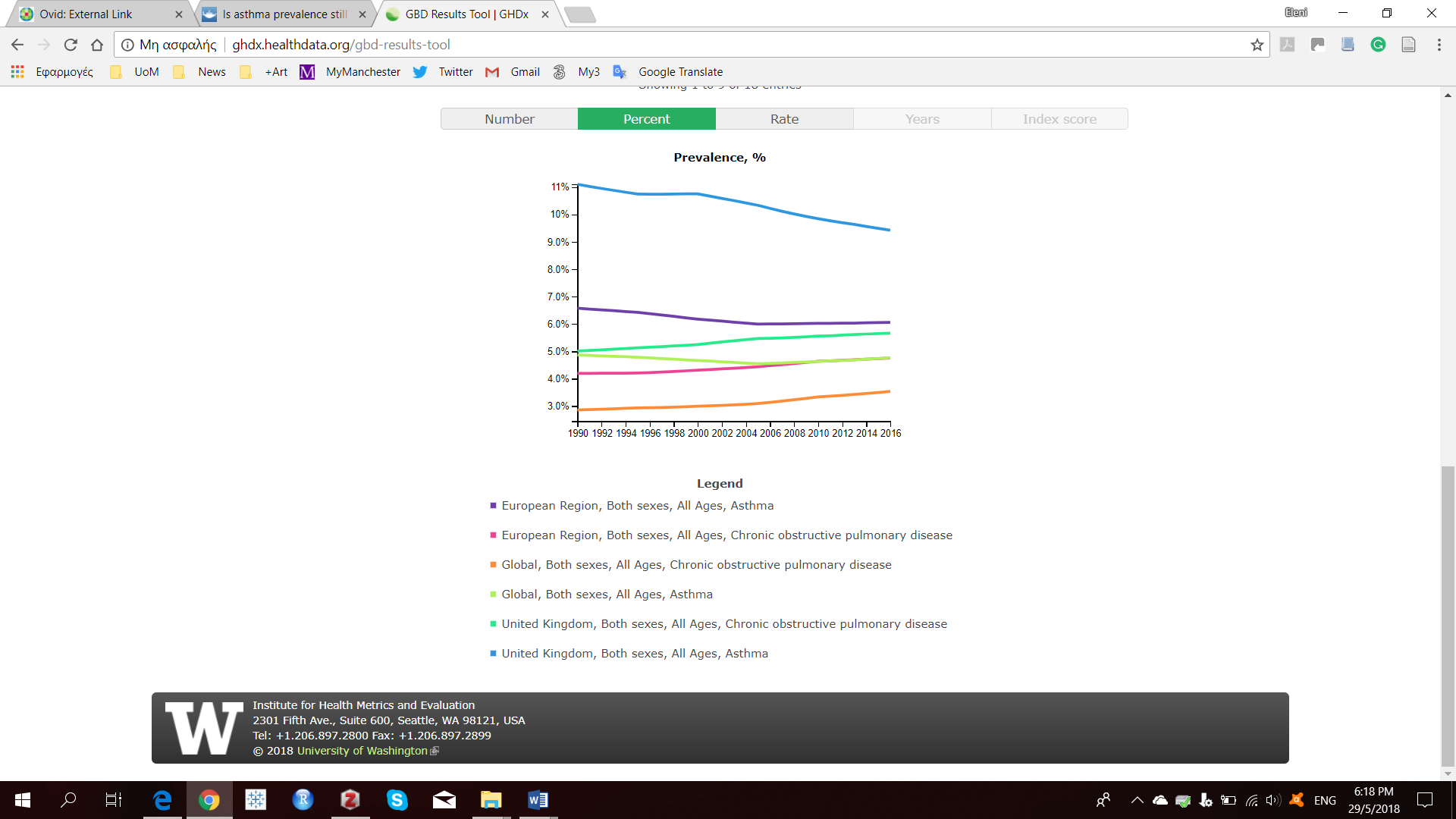
**Figure 1**: Infographic showing good and bad habits for COPD patients (Dohyung 2017)



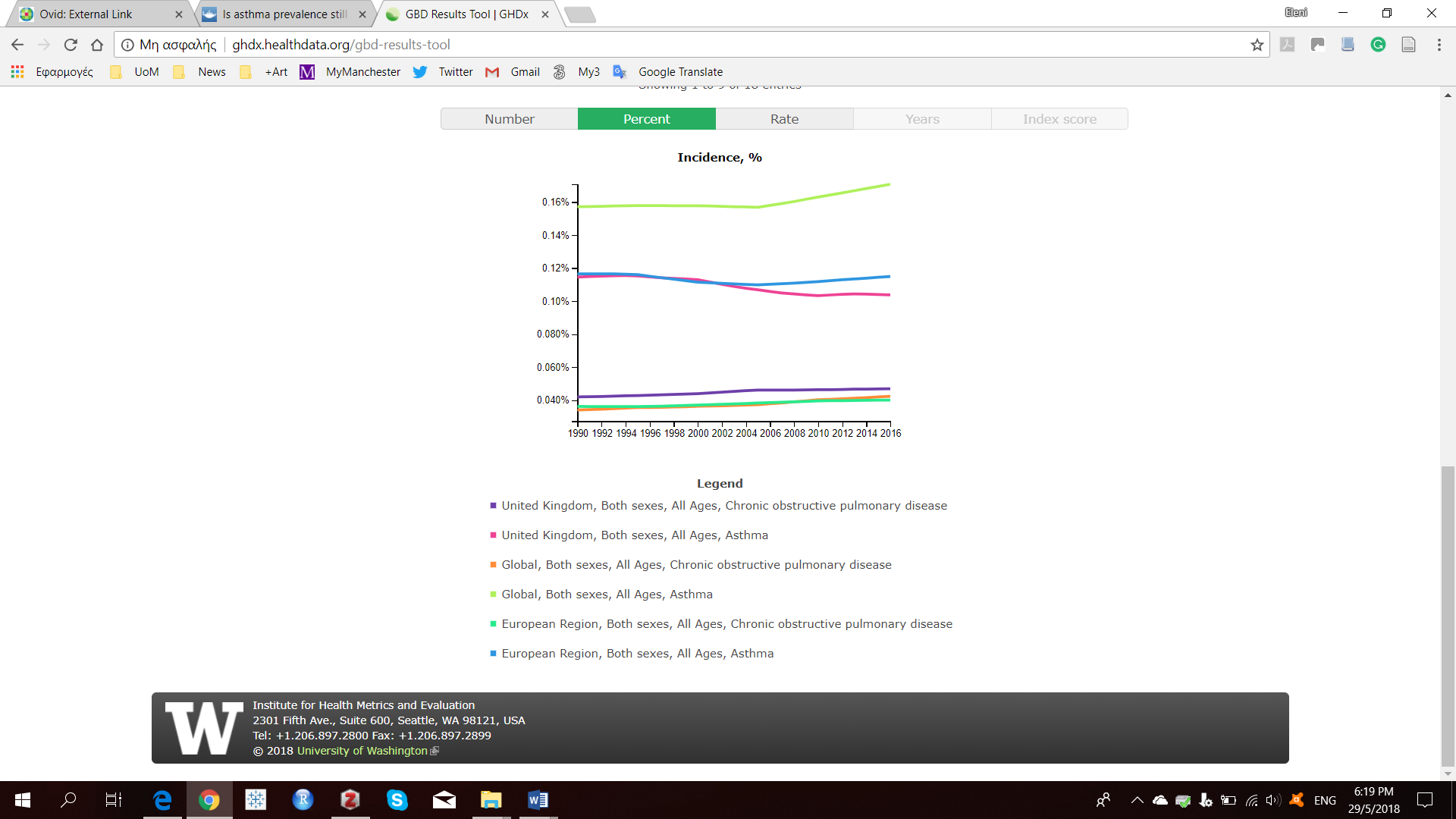
**Figure 2:** People with asthma who have received either emergency or unplanned care at a hospital or out-of-hours centre during 2017 (Asthma-UK 2017).



**Figure 3:** Further break down of the above figure; proportion of people with controlled and uncontrolled asthma who have received hospital care in 2017 (Asthma-UK 2017).



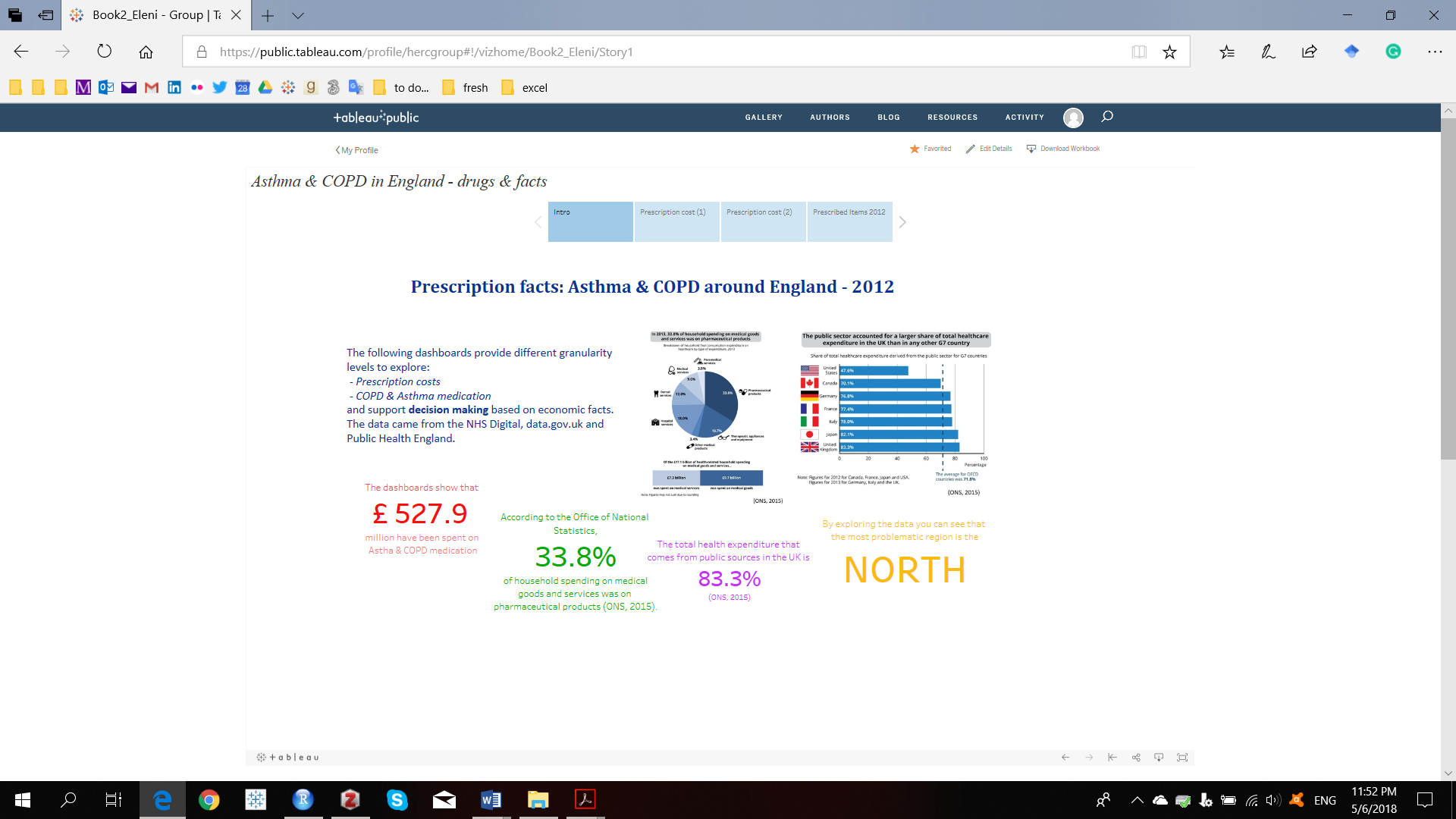
**Figure 4:** Prevalence (%) of asthma and COPD in the UK and EU (GHDx 2016)



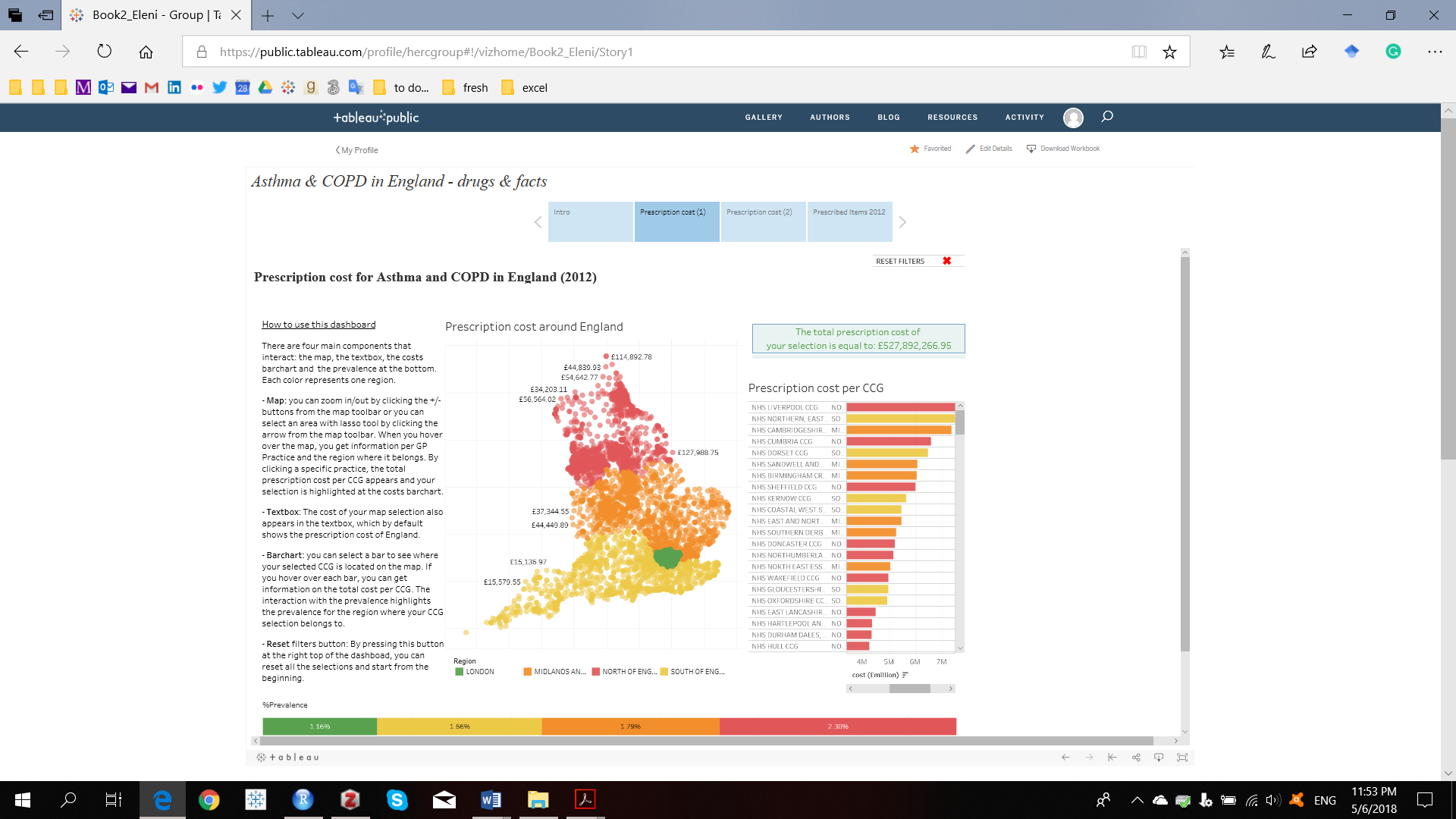
**Figure 5:** Incidence (%) of asthma and COPD in the UK and EU (GHDx 2016)

**Figure 6:** Five main words that describe data quality and additional terminology (Weiskopf and Weng 2013)

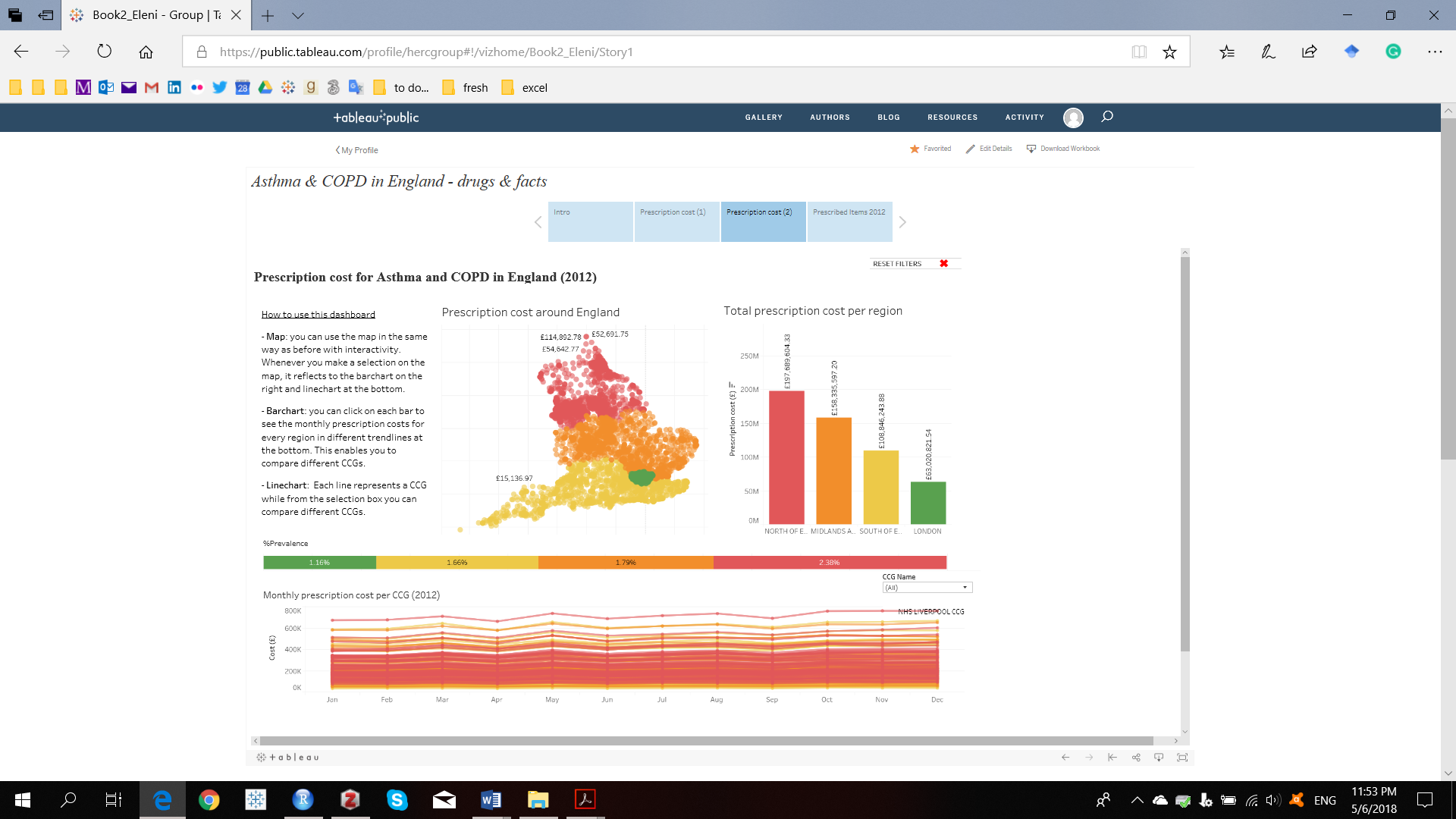
**Figure 7:** UK health data landscape; \*NACAP programme is currently in progress (McNeil 2017; NACAP 2016; McDonnell et al. 2017)



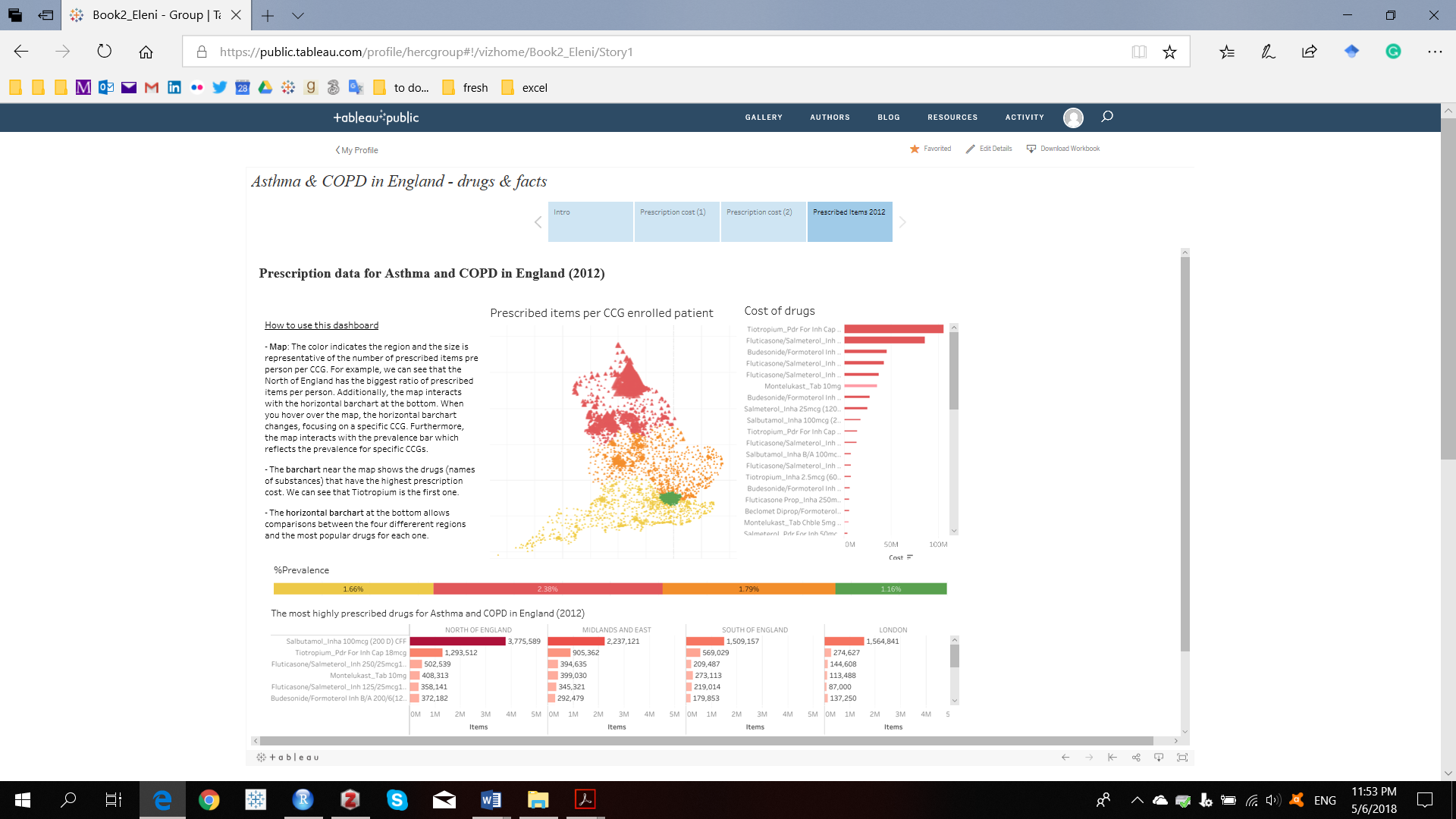
**Figure 8:** An introduction page explains main ideas and facts that the next dashboards reveal on prescription costs



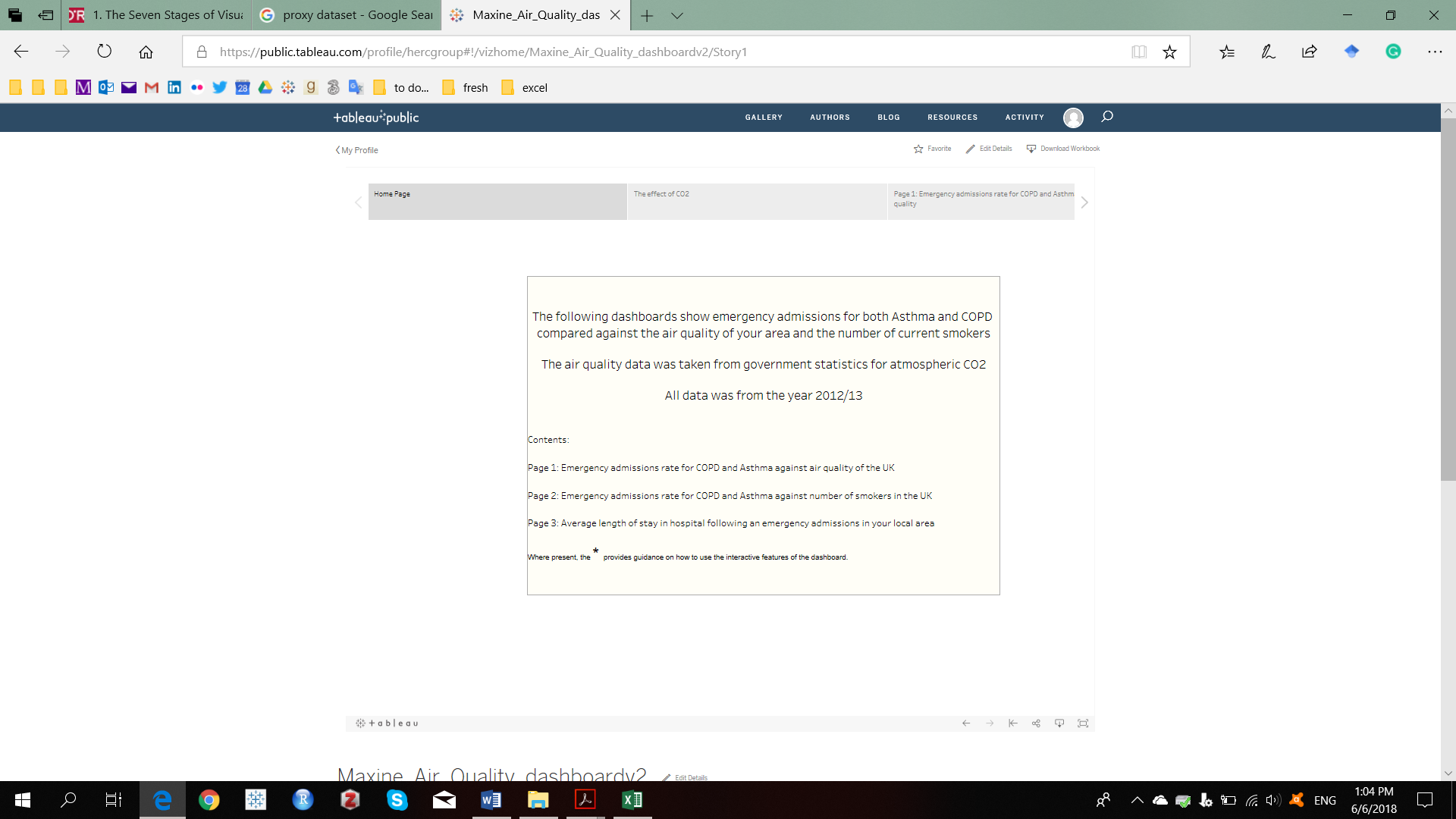
**Figure 9:** This dashboard emphasizes on the north, showing that the biggest amount of money is spent in Liverpool and that the costs follow the prevalence trend (interactivity has been included).



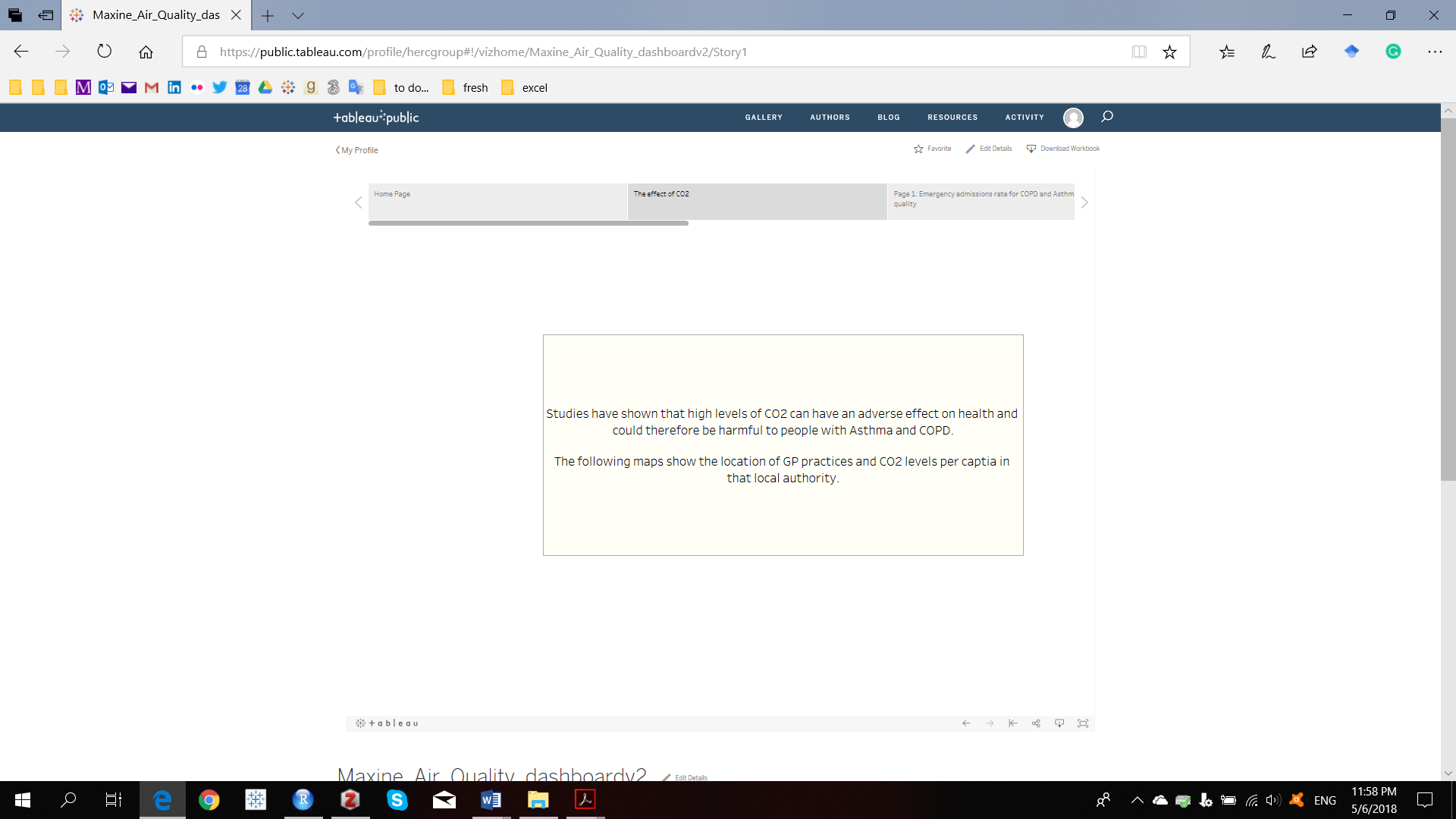
**Figure 10:** This dashboard is an extension of the previous, giving the opportunity for the user to make comparisons between different CCGs and to check monthly prescription cost trends for 2012 (interactivity has been included).



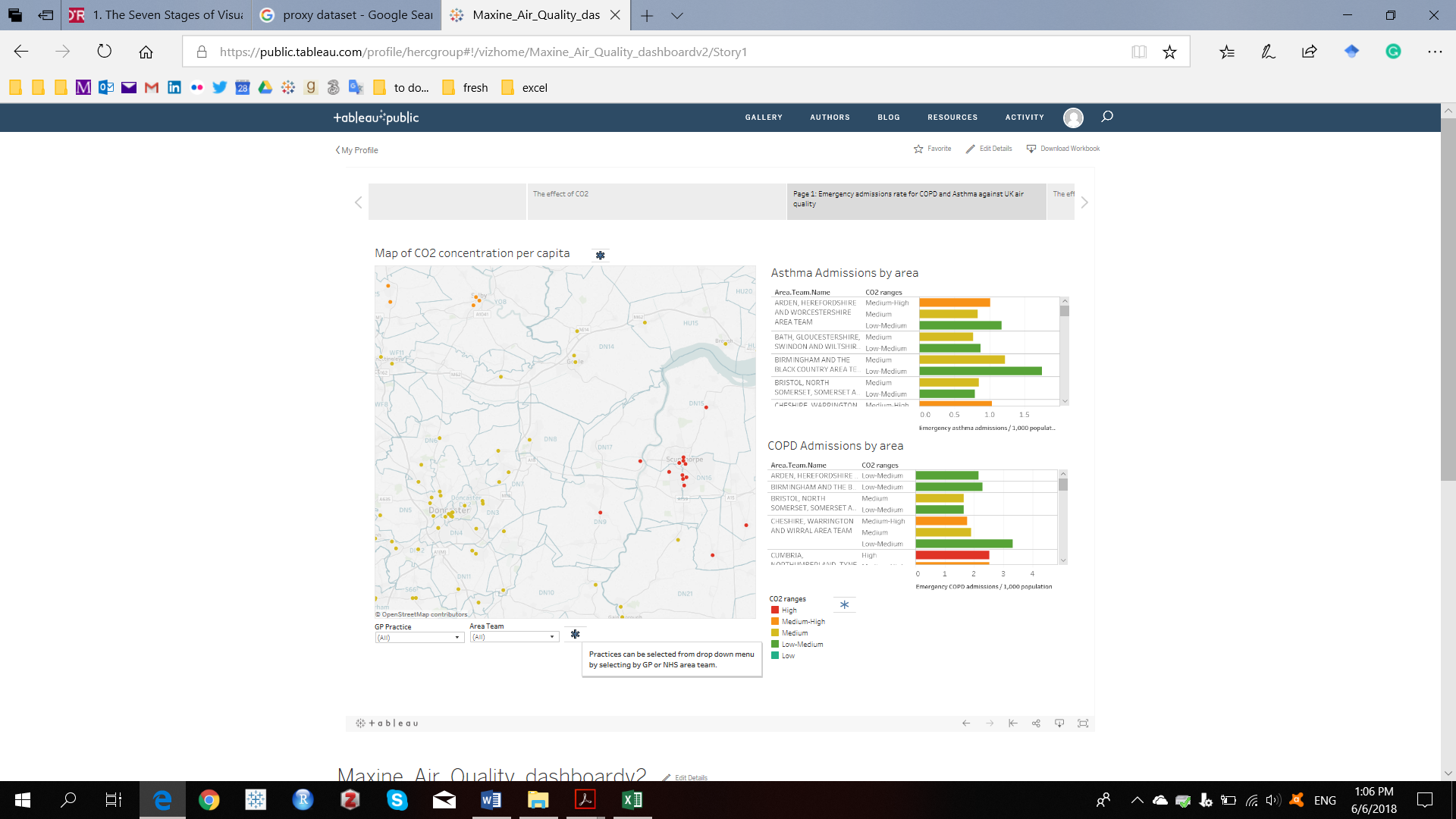
**Figure 11:** This dashboard shows the number of prescribed items per CCG (interactivity has been included).



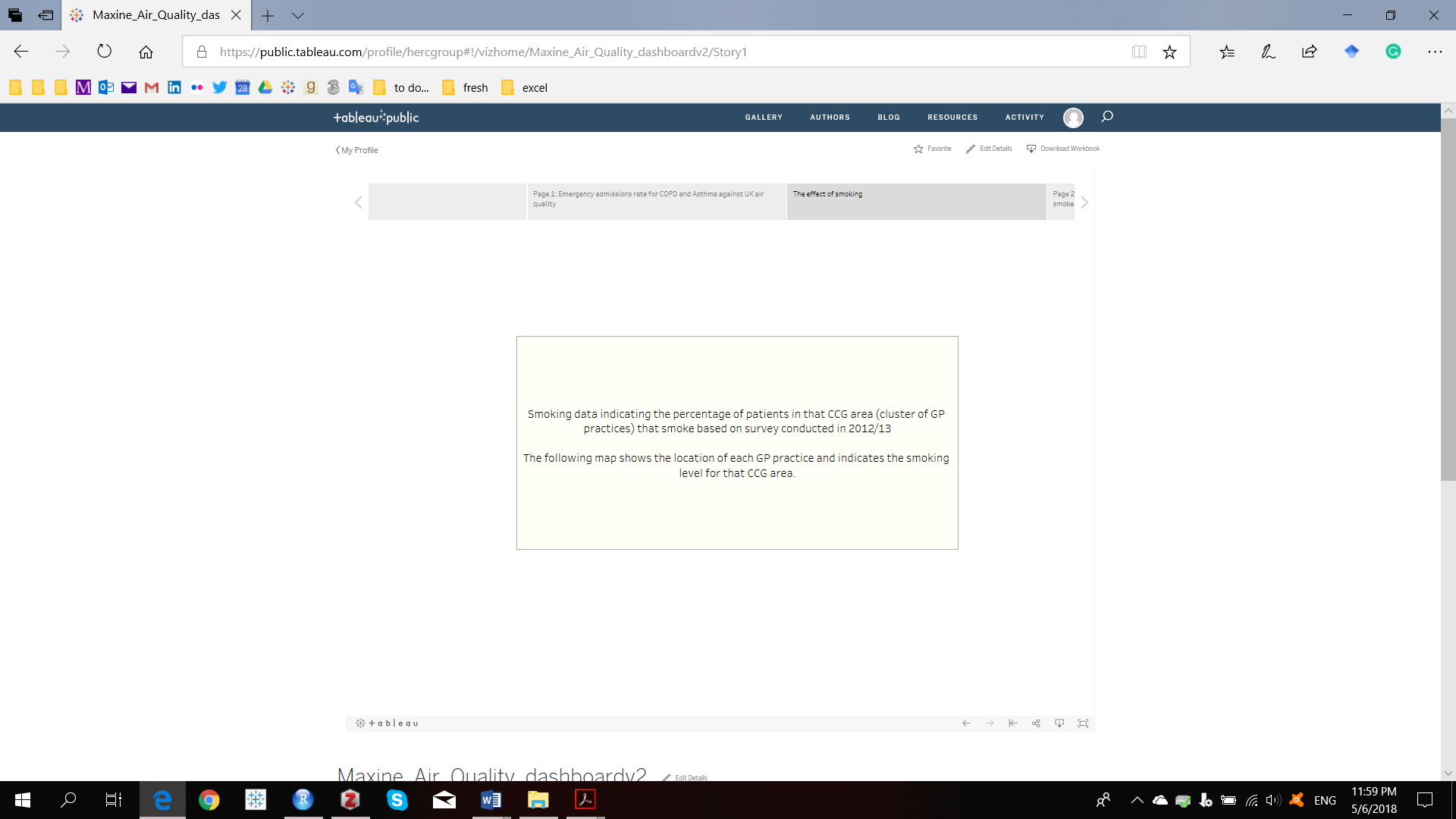
**Figure 12:** Introduction dashboard



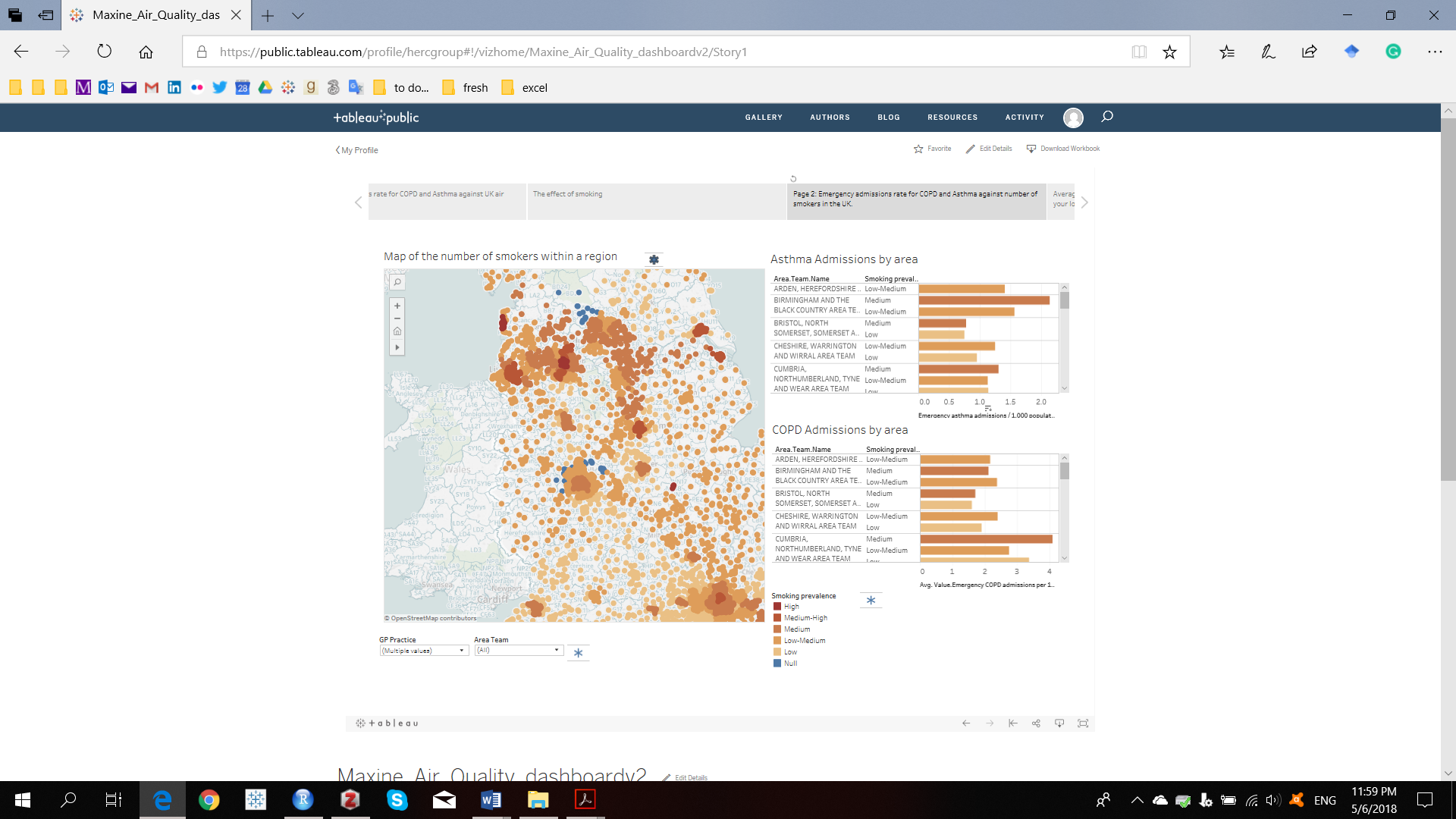
**Figure 13:** Introduction dashboard



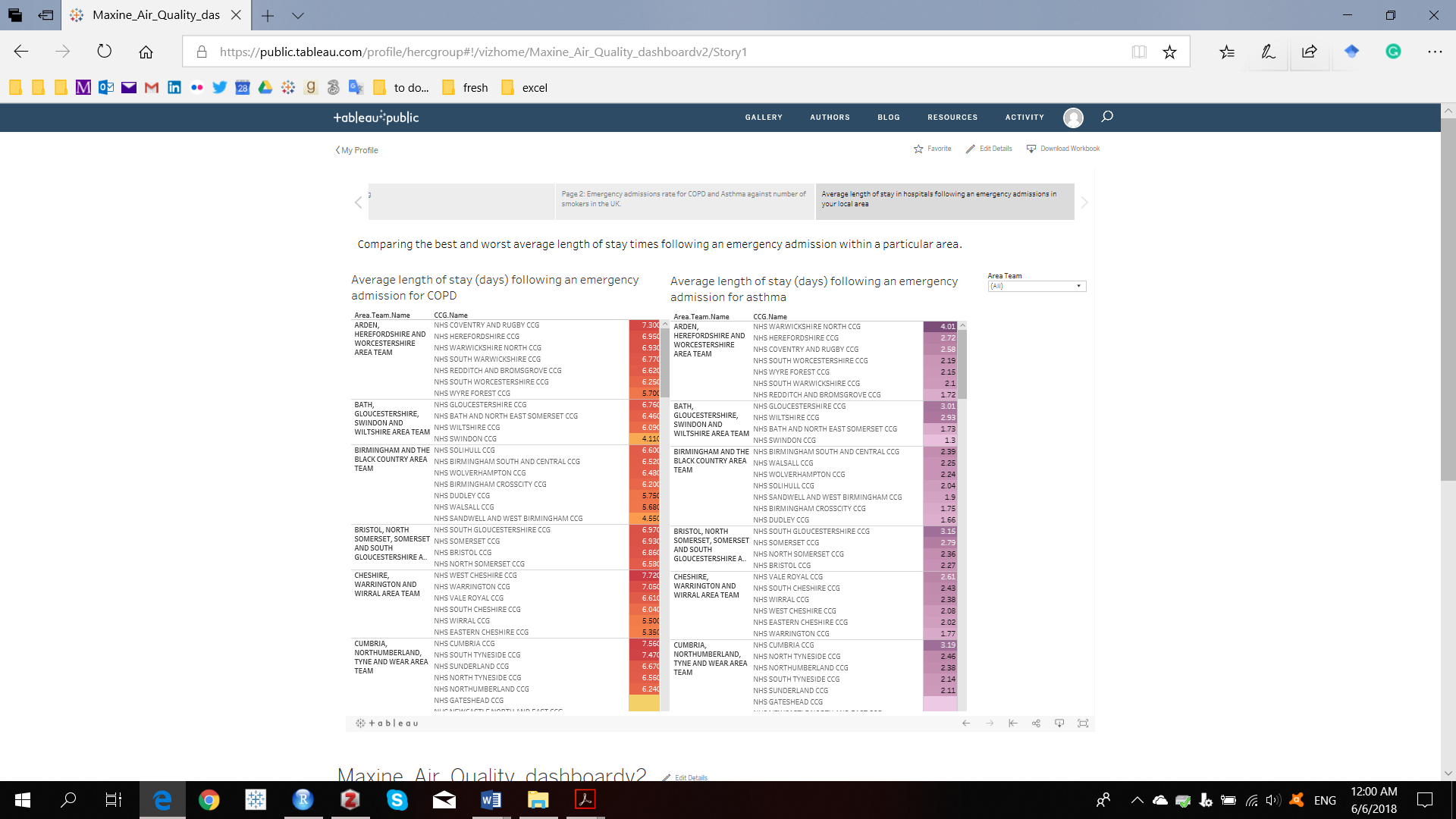
**Figure 14:** In this dashboard we can explore different regions to check carbon dioxide emissions and any potential relationship between hospital admissions for asthma and COPD (we hover over one star to show an example of its functionality.



**Figure 15:** Introduction dashboard



**Figure 16:** In this dashboard we can explore different regions to check how smoking prevalence affects hospital admissions for asthma and COPD.



**Figure 17:** Average length of hospitalisation in days in case of an emergency for both asthma and COPD.

1. PHE has been established since 2013 as an Executive Agency of the Department of Health [↑](#footnote-ref-1)
2. Available at <https://qof.digital.nhs.uk/> [↑](#footnote-ref-2)
3. <https://www.gov.uk/government/statistics/uk-local-authority-and-regional-carbon-dioxide-emissions-national-statistics-2005-2013> [↑](#footnote-ref-3)
4. <https://data.gov.uk/dataset/176ae264-2484-4afe-a297-d51798eb8228/gp-practice-prescribing-data-presentation-level> [↑](#footnote-ref-4)
5. Twelve different datasets – one for each month – have been linked together. [↑](#footnote-ref-5)
6. A dataset that is a linkage between data from PHE, NHS Digital and the Ordnance survey. We refer to this dataset using the word “initial” dataset. [↑](#footnote-ref-6)
7. <https://digital.nhs.uk/data-and-information/publications/statistical/patients-registered-at-a-gp-practice/april-2013> [↑](#footnote-ref-7)