The devil, the detail, and the data

Andrew D Garrett

Address for Correspondence: Andrew Garrett. ICON Clinical Research, 500 South Oak Way, Green Park, Reading, RG2 6AD. UK

Email: andrew.garrett@iconplc.com

Abstract

Digitalisation has created a world awash with data from sources almost unimaginable 50 years ago. Using recent examples from climate change, the Covid-19 pandemic, official statistics, and artificial intelligence, the address will illustrate the importance of understanding the data generating process and the data about the data. It will be argued that statisticians and other data professionals have an increasingly important role to play as advocates for well-defined and well documented data generation – particularly as the distinction between inputs and outputs becomes blurred. For the Society this does not represent a shift away from statistics, rather it re-establishes the link to our roots.

Keywords: AI, climate change, covid-19, data generating process, meta data, official statistics.

[The address of the President delivered to the Royal Statistical Society on Wednesday, June 26th, 2024.]

1. Introduction

I am reliably informed that inaugural addresses first started around 1869 and presidential addresses later so my attempts to uncover what Lord Stanley might have thought on becoming the 13th president of the Society in 1857 were sadly thwarted. So why look to Lord Stanley? Well, you must start somewhere when preparing your own address and list of presidents dating back to 1834 is somewhat overwhelming and mightily humbling. A random sample of previous presidential addresses might have been appropriate, but instead I looked for connections. More on those connections later.

As David Hand pointed out in his presidential address "statistics is so ubiquitous" such that the learning curve for a president is steep (Hand, 2009). Whilst true, as David also notes, this also presents the opportunity to represent a society that plays everywhere. If Statistics is ubiquitous then so is data and in this address my intention is to take this broad view of both statistics and data. Indeed, the aim is to place the spotlight on the data – in particular, how the data is generated (the data generating process) and more broadly the data about the data (the meta data). Although these topics inevitably have been covered extensively before, emerging challenges bring twists and turns, and in some cases retreats.

Peter Diggle in his presidential address (Diggle, 2015) took on the topic of data science – the threats and the opportunities – and almost 10 years on, it is difficult to speak about data without speaking about the impact of data science and artificial intelligence. A landmark year for generative AI (large language models), 2023 was the year that Pandora's box opened with concerns of existential threat and the emergence of regulation.

Sylvia Richardson in her presidential address (Richardson, 2022) noted that: “Data never exists without relation to something, and the method used to collect data needs to be scrutinised through the statistical magnifying glass, with close attention paid to the nature, representative or not, of the selection process by which the data is collected”. This address also presents an opportunity to consider data “in relation to something”. To showcase some areas where the Society has been particularly active of late – these include climate change, the Covid-19 pandemic, official statistics, and AI– and to use these to illustrate the importance of understanding the data generating process and the meta data.

In David Hand’s address 15 years ago (Hand, 2009), he was prescient in terms of noting statistics “extraordinary ubiquity in modern life” and where awareness of statistics was needed. He highlighted climate change and epidemiology (such as respiratory disease) and asked: “how to raise public awareness of the key role of statistics in all these and other activities”. In John Pullinger’s Presidential address (Pullinger, 2013) he lamented the results of a public poll that pointed to “limited awareness or attention given to statistics” and his hope to turn things around by 2020 such that there would be “widespread statistical awareness” and that “Statistics will be valued as useful to people”. He continued: “Greater understanding will improve trust. Lack of statistical skills will be perceived as a disadvantage”. John was, of course, referring to the excellent work that many were doing including the GetStats campaign launched in 2010. However, in 2020 the Covid-19 pandemic struck, and the media (including social media) were awash with data, graphs, analyses and interpretation. As Sir Ian Diamond, the UK’s national statistician, in an interview for the Society’s Real World Data Science Platform (Tarren, 2023a). noted “one of the things to come out of this dreadful pandemic was that people across the country became more data literate, and more demanding of the data, and more able to interpret data”. Sylvia Richardson (Richardson, 2022), whose presidency coincided with the pandemic noted: “The awareness that statistics, their production and their interpretation have a direct impact on everyone’s life has been heightened in the public consciousness by the Covid-19 existential threat.”

This leads to thoughts about the relationship the Society has with data and those that work with data. It also links to developments around the Society’s five-year strategic plan (2024-2029) and to our Royal Charter that details “promoting the public understanding of statistics and the competent use and interpretation of statistics.” It also links to matters of trust and to how we interact with others. But the devil is in the detail and the detail starts with the data.

2. The data generating process

Data is ubiquitous and it is generated in many ways. This data generating process has important implications for extracting information from the data and for drawing reliable conclusions from it. There are however two stand-out methods for generating it that enable statisticians to make statements that address bias and uncertainty. These are randomisation and random sampling, that will be discussed in sections 2.1 and 2.2 respectively. Some other data generating processes will then be discussed in sections 2.3 to 2.8.

2.1. Randomisation

Past President Ronald Fisher developed the randomised experiment between the World Wars, and it was another past President, Austin Bradford Hill who is credited with introducing the concept to clinical research in the late 1940's (Medical Research Council, 1948). It is the instrument of randomisation that ensures that experimental and control interventions are allocated without bias and provides the probabilistic basis for comparisons between different

interventions – forming the basis for causal inference through the control for potential confounding factors (Breslow, 2001). Randomised and controlled trials (RCTs) are well known for their use in clinical trials, but they are used in many other research areas including agriculture (e.g. evaluating crop yields), animal research and in social science.

Randomisation has three important functions. It provides protection against selection bias in the assignment of an intervention to experimental units (for instance treatments to patients in a clinical trial). Over all randomisations, it generates intervention groups that are balanced with respect to factors - known and unknown, measured and not measured - that influence outcome but are independent of the intervention assignment. Finally, given that the randomisation procedure was not violated, it enables test statistics to be generated for intervention comparisons (Garrett, 2006).

Although simple in concept randomisation can be applied in various ways including to treatment sequences (cross-over designs) in clinical trials and to groups of subjects (cluster designs). It can also be incorporated within more complex designs that include blocks and strata originating from its agricultural research origins. Indeed, the design links to the subsequent analysis such that the model selected should reflect the features of the design – that is, it should reflect the data generating process. The short-hand version is often quoted as *analyse as you design* to produce valid inferences. Analyse a cross-over design as a simple parallel group one and you run into problems in estimating the correct standard error for the treatment comparison (Senn, 2015)

Randomisation is closely associated with drug development, which together with the blinding of investigators and subjects to randomised treatment, underpins the requirement to provide substantial evidence in the form of “adequate and well controlled investigations” (FDA, 1997) to support regulatory approval. Examples include the SARS-CoV-2 treatment and vaccine trials described in Section 3.2.

2.2 Random sampling

Random sampling is a probability sampling technique that requires a sampling frame. The straightforward concept is to ensure that every member of the population has a known (and usually equal) probability of being included in the sample (Thomas, 1977). Furthermore, every sample of a given size (with k members, say) has an equal probability of being selected. An important consideration, notably for small populations, is sampling without replacement – that ensures that members of the population cannot be selected more than once. At its heart is the aim to obtain an unbiased estimate of some characteristic of a population and to design the sampling scheme such that the corresponding variance of the of the estimate (the standard error) is sufficiently small to meet the needs of the research. The REACT-1 survey (Elliott et al, 2023) that was used to estimate SARS-CoV-2 infection prevalence in England is a classic example of a random sample and is discussed in more detail in Section 3.2. Importantly a population is not restricted to people and can include amongst other things businesses (to produce economic statistics) and products on a manufacturing production line (to estimate defects).

Like randomisation, variations in random sampling exist to accommodate different needs including stratified random sampling where the population is divided into strata or groups based on specific characteristics. Past president Sir Arthur Bowley undertook fundamental work in this area although it is Neyman (Neyman,1934) who is described as laying the foundations of probability sampling. Random sampling is an extensively researched area with challenges such as missing data and small area estimation (Rao and Fuller, 2017).

Since both randomisation and random sampling bring control over the data generating process, they also provide an opportunity control for bias in terms of pre-specification – the ability to match the proposed statistical analysis to the data generating process before the data is generated. This avoids the likelihood of cherry-picking the most favourable results - a term often referred to as p-hacking.

2.3 Observational data

Observational data is a broad term, since it simply implies data that is observed. In some sense it is an umbrella term. For research purposes, there is an important distinction between data that is generated prospectively according to some research protocol, including with the requirement for individual consent if studying humans, versus data that is retrospectively accessed, in de-identified form if studying humans using various legal gateways to access data. Prospectively generated data may include clinical trials with no randomised control group, but in which the data is generated in an almost identical way as for a randomised controlled trial, with set study visits where defined procedures are undertaken. In this sense these studies are interventional. Commonly used examples are long term extension clinical trials in drug development where once patients complete a randomised trial, they may enter an open-label extension protocol where all patients receive the same treatment (usually to gather long term safety data on that treatment).

Observational data is common in epidemiology and social science research. Examples of studies that collect data prospectively, longitudinally but in a non-interventional manner are Our Future Health (2024) that aims to recruit 5 million volunteers in the UK to research disease prevention, detection, and treatment and UK Biobank (2024) that has access to data from half a million volunteers via de-identified medical and genetic data. In the social sciences examples include the National Child Development Study (NCDS, 2024) which follows an initial cohort of 17,415 people born in a single week in 1958 (in England, Scotland, and Wales) – collecting data on their physical and educational development, economic circumstances, employment, attitudes etc.

2.4 The Census

A census is a unique data generating process that aims to prospectively capture data on all persons in a country on a specific day of the year and is conducted at set-intervals through time. In the case of the UK, the first census was conducted in 1801 and with a few exceptions (e.g. 1941 during the second world war) has been conducted every 10 years since and most recently in 2021 (2022 in Scotland). The census is compulsory, and fines can be imposed for non-completion. In England and Wales, the Office for National Statistics (ONS) is responsible for planning, conducting, and reporting the census with all information anonymised and retained securely for 100 years. The ONS (ONS, 2022) describes it as follows:

“The census asks questions about you, your household and your home. In doing so, it helps to build a detailed snapshot of our society. Information from the census helps the government and local authorities to plan and fund local services, such as education, doctors' surgeries and roads.”

2.5 Administrative data

Administrative Data Research UK (ADR UK, 2024) defines administrative data as “information created when people interact with public services, such as schools, the NHS, the courts or the benefits system, and collated by government”. This data is usually collected as part of the service being offered – to be able to function and deliver a service effectively

and efficiently, rather than to answer research questions (Hand, 2018). They reflect the increasing digitalisation of services and the two-way digital interaction with government. Although these data can give the impression of being a complete data set (Data = ALL), there is less control over the data generating process and the importance, or lack of, placed on specific data by those that complete or collect it may not be immediately apparent to the statistician or data professional. Data may be collected (for example, a government form completed) in an order that does not correspond to that expected by a statistician and what is considered important may be viewed through a different lens. This can result in important data for research purposes being incomplete, missing or completed in a manner without much thought to accuracy (Garrett, 2016). The data required to answer a specific research question may be very different from the data needed to provide a service and legacy systems may limit the potential to add or modify variables that would produce better statistics (Garrett, 2018).

2.6 Smart Data

Smart Data Research UK (Smart Data UK, 2024) defines smart data as “the data generated through engagement with digital systems, devices and sensors” and these include “mobile apps, navigation systems, social media, sensors in consumer devices and the environment and digital transactions”. Smart data often reflect the passive collection of data and in the context of social media the use of this data by others may represent the price to be paid for free access. Vichi and Hand (2019) discuss the challenges of extracting trusted and reliable aggregate information from these data sources and make recommendations to the producers of these smart data.

2.7 Simulated or synthetic data

Simulated data is computer generated data that is created using various assumptions. It is often used by statisticians to compare the performance of various statistical tests or estimation methods and it can be used to help design studies by understanding whether a study is likely to meet its objectives under a range of assumptions – ranging from the plausible to the extreme. Typically, the data is generated randomly from probability distributions using a specified model. Synthetic data can be viewed as being based on authentic or real data that is changed to a sufficient extent such that the original individual data (to retain privacy and confidentiality) is unidentifiable, whilst retaining the original properties of the data set. The concept here is to use data that is more robust and consequently provide more reliable results compared to alternatives. As will be discussed in section 3.4, with increased digitalisation of outputs and advances in generative AI, the distinction between inputs and outputs is beginning to become somewhat blurred particularly with what is termed, hallucinogenic AI. In this respect, inputs become outputs that become inputs.

2.8 Meta data

Meta data is the data about the data. It can include details such as who or what created the data, when (a date and time stamp) and where. In the context of social media, it could include the location and duration of an interaction with a smart phone. It includes information about the structure of the data too – the variable name, the label, the format etc. For derived data, it describes how the data is derived - from which data, using what formula, including how missing or partially missing data (dates, for instance) is handled. For data linkage, the meta data describe how different data sets are combined – joined, merged or linked.

Meta data includes information on where the data is stored that enable data to be accessed and retrieved over time.

Importantly meta data provides traceability, which can be important for quality, transparency and hence trust. It provides the roadmap from the start of the data journey to the end. It enables data sets and analyses to be reproduced – another factor in building trust as it enables other researchers to conduct their own independent analyses of the same data – identify differences and determine how sensitive the results are to the decision made regarding, for instance, the handling of missing data. Meta data also allow teams of statisticians, programmers, and data professionals to work on the same project, including conducting quantity control of the work conducted by others by checking that the programming is consistent with the specifications – the meta data – that the specifications are consistent with adjacent or related work (for instance, with other clinical trials in the same programme of work). It provides the detail, and all too often the devil is in the detail.

3. Some examples

In the following four sub-sections, I have selected four topical areas and use these to illustrate the importance of the data generating process and of the meta data associated with the outputs. The examples include a mix of data that produce statistics and statistics that become data.

These four areas are: climate change, the SARS-CoV-2 pandemic, official statistics, and artificial intelligence.

3.1 Climate change

In her presidential address, Deborah Ashby pointed out that “An area that our founders did not consider is the environment”, noting the “the long-standing contribution of statistics and statisticians in this area, which is becoming even more pressing with rapid climate change”. (Ashby, 2019).

Given that the Intergovernmental Panel on Climate Change (IPCC) Special Report on the impacts of global warming (IPCC, 2018) refers to the period 1850-1900 as the pre-industrial baseline, it is perhaps not surprising that a society founded in 1834 did not anticipate the effect of emitting gases including CO_2 and methane into the atmosphere at scale. The world was very different then. That is not to say that concerns were not raised early. The first use of the term greenhouse effect is attributed to the Swedish scientist Svante Arrhenius in 1896 (Arrhenius, 1896) who estimated the effect of heat absorbing gases on mean ground temperature. In 1912, the New Zealand Rodney and Otamatea Times, Waitemata and Kaipara Gazette included the headline “Coal consumption affecting Climate” which was taken from the March edition of *Popular Mechanics* magazine (Molena,1912). The text states:

“The furnaces of the world are now burning about 2,000,000,000 tons of coal a year. When this is burned, uniting with oxygen, it adds about 7,000,000,000 tons of carbon dioxide to the atmosphere yearly. This tends to make the air a more effective blanket for the earth and to raise its temperature. The effect may be considerable in a few centuries.”

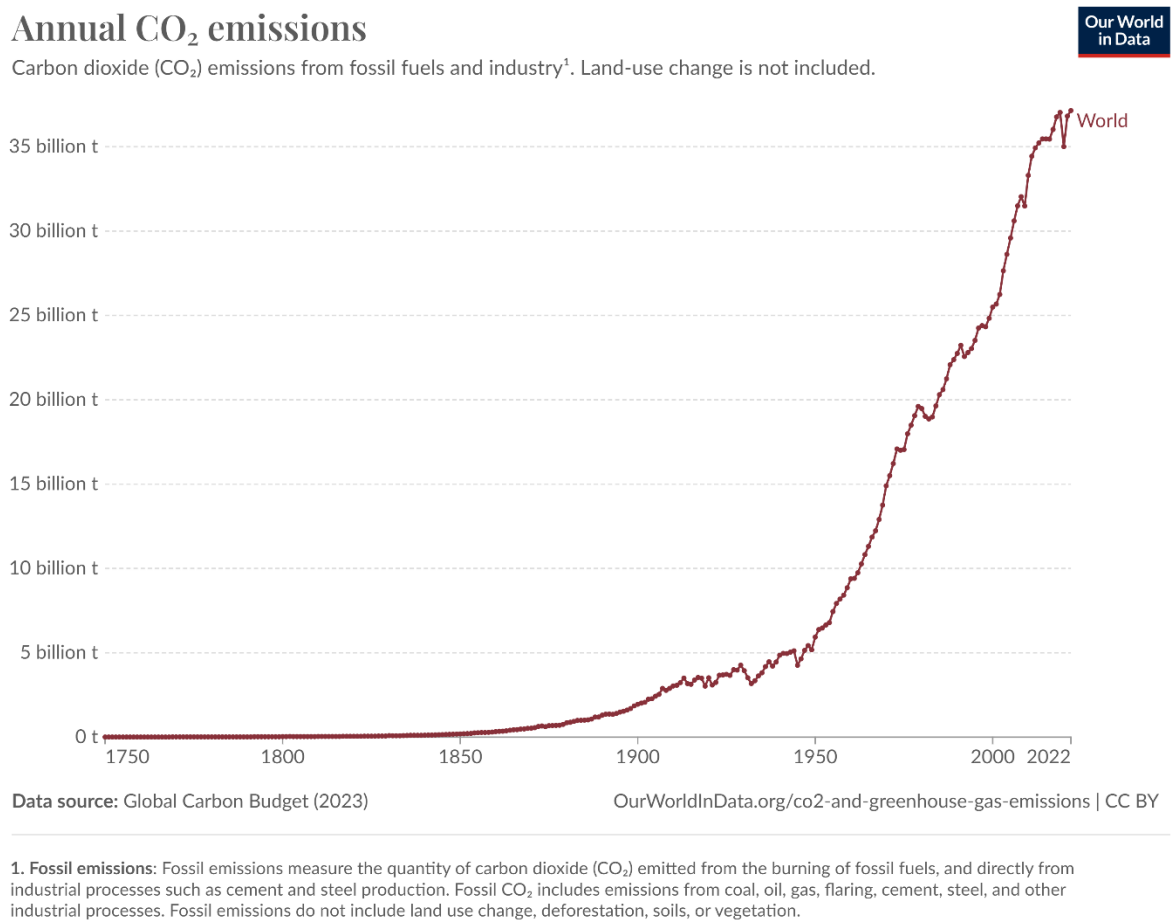
Limiting global warming to 1.5°C was first mentioned in the Cancun Agreement in 2010 (United Nations Climate Change Conference, Conference of the Parties (COP)16) while the Paris agreement in 2015 (COP21) formally adopted 1.5°C as a limit. That is ‘holding the increase in the global average temperature to well below 2°C above pre-industrial levels and pursuing efforts to limit the temperature increase to 1.5°C above pre-industrial levels’. The temperature limit has since been restated as a ‘defence line’ or ‘buffer zone’, instead of a ‘guardrail’ up to which all would be safe’. (IPCC, 2018).

The 1.5°C target is now well established and commonly referred to, but is it well defined and understood in terms of how it is calculated? This has been an area that the Society's Climate Change Task Force (RSS, 2024) has delved deeper into to understand the data about the data.

Global surface temperatures are estimated using a combination of land and sea surface temperatures from multiple locations. Land surface temperature (LST) measurements are taken at fixed locations whereas sea surface temperatures (SST) are now taken most often from buoys, but also still from ships. Historically LSTs have corresponded to where people have lived or worked and SSTs from where they have sailed – the latter affected by two world wars that saw shipping severely restricted – and geographic coverage has increased over time. Adjustment for non-climatic artefacts is termed homogenization – something that impacts SSTs more than LSTs. Fewer observations are required to estimate the global average when these observations are geographically dispersed as observations closer together tend to be more similar (Thorne, 2016). There are four major data sets that are used to study global temperatures over time and according to the National Aeronautics and Space Administration (NASA, 2021) "Today's temperature data come from many sources, including more than 32,000 land weather stations, weather balloons, radar, ships and buoys, satellites, and volunteer weather watchers".

As stated earlier, the pre-industrial baseline that is commonly (although not uniformly) used is the one stated in the IPCC report – that is 1850-1900. It appears to represent a period where a sufficient number and geographic spread of temperature measurements were available encompassing both LST and SST, whilst pre-dating the major increase in human-induced greenhouse gas emissions. The industrial revolution in Great Britain began a century earlier in the middle decades of the 18th century although as Figure 1 demonstrates (Our Word in Data, 2024), global CO_2 emissions due to fossil fuels and industry increased materially much later. However, it is not the only baseline that is used to report temperature anomalies. NASA (2021) uses 1951-1980 while the US's National Oceanic and Atmospheric Administration's (NOAA, 2024) Earth Observatory uses 1901-2000. In Europe, the Copernicus Climate Change Service (C3S, 2021) implemented on behalf of the European Commission uses the IPCC baseline but also 1991-2020 as a reference period.

Figure 1. World annual CO_2 emissions (1850-2022)

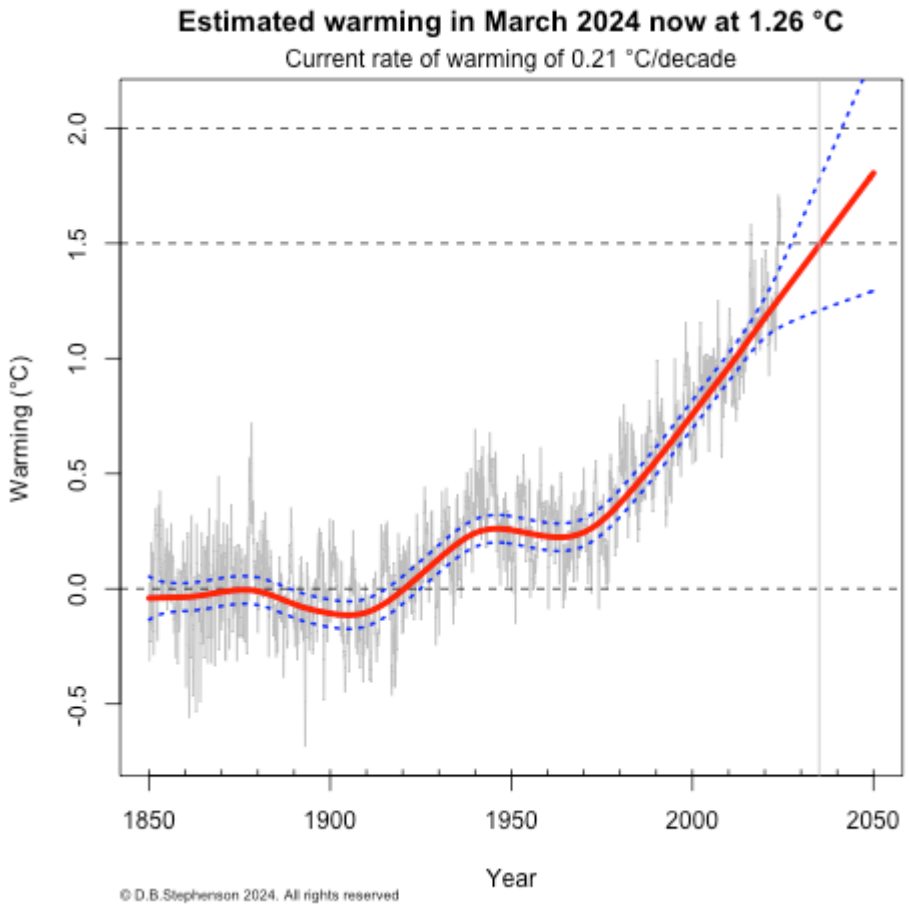


Once the baseline is defined, then the next step is to determine how a change in the global mean temperature is defined and then estimated. As Betts *et al* (2023) states "It might come as a surprise, then, to hear that the Paris statement contains no formally agreed way of defining the present level of global warming". Global temperatures vary naturally over both short and long periods and are impacted by well researched phenomena such as El Niño, so fluctuations are to be expected. To account for these fluctuations the IPCC 6th assessment report (IPCC, 2023) chose to smooth the data using a 20-year moving average, such that if the moving average exceeded 1.5 degrees, then the mid-point of the 20-year period would be the year the limit was exceeded. This definition has limitations – notably it is a lagging measure of change since once reached, the year that 1.5 degrees was breached was 10 years earlier. In other words, we will not know what the moving average is for 2024 for another 10 years. Strictly since 20 is an even number, the middle value of the moving average is the mean of two years, not one. Overall, it represents risk around slow detection, confused communication, and delayed action.

Betts et al (2023) state that "A more instantaneous indicator of the current level of long-term warming is needed" something that the Society's Climate Change Task Force also concluded. Suggestions include: "finding the end point of a linear trend over the past 30 years, using more sophisticated methods for statistical smoothing over short time frames; and calculating the human contribution to warming from data on changes in the concentrations of greenhouse gases and aerosols." The Climate Change Task Force co-chair David Stephenson, estimates monthly the long-term trend (with 95% confidence interval) by generalised additive modelling (Wood, 2006) of the anomaly values as the sum

of a smooth cyclic function of year (the trend), a smooth function of calendar month, and an irregular [autoregressive] AR(2) noise term having a month-to-month persistence (Stephenson, 2024). This is illustrated in Figure 2.

Figure 2. Long-term trend estimation of global mean temperature change with 95% confidence intervals together with monthly data (baseline 1850-1900)



Climate change and sustainability is a complex and multifactorial topic that is awash with data. Interpretation and communication are consequently challenging and nuanced. However, for headline data pertaining to global average temperature increases that are so widely publicised, discussed and dissected it is truly remiss how a 1.5°C target has taken so long to be clearly defined and communicated – and arguably how poorly it is defined statistically. The impact of the devil's furnace has been hidden away for too long.

3.2 Covid-19 pandemic

The SARS-CoV-2 pandemic was undoubtedly a big data pandemic with data summarised daily and presented to the public via various news channel usually in graphical form. Early on in 2020 there were data on infections, hospitalisations and deaths with discussion on how these were impacted by various non pharmaceutical interventions (NPI) including, *in extremis*, lockdowns. Following the unprecedented speed of vaccine development, daily updates were supplemented in 2021 with data on vaccinations. My predecessor as President, Sylvia Richardson, eloquently describes in her presidential address how the pandemic dominated and changed the course of her presidency - notably through her co-

chairing (with past president Sir David Spiegelhalter) of the Society's Covid-19 Task Force. She described how statisticians and data scientists needed to adapt to a world where agility and responsiveness were key. She noted how "Interesting challenges to conventional statistical practice arose during the pandemic from the need to analyse in real-time, messy data from diverse sources to answer constantly changing questions from the health authorities (Richardson, 2022).

However, reporting the number of deaths by date and cause, proved far from straightforward and it took a few months for the meta data to reveal themselves subsequently leading to important revisions. For instance, it became apparent that a death with a laboratory-confirmed positive COVID-19 polymerase chain reaction (PCR) test was originally being counted regardless of time since test – less of an issue in the early weeks and months of the pandemic, but as time progressed and PCR testing increased substantially this definition began to include deaths that were clearly unrelated to these much earlier positive test results.

Public Health England (subsequently replaced by the UK Health Security Agency) clarified and modified the definition on 12th August 2020 (UKSA, 2022). This narrower measure became "A death in person with a laboratory-confirmed positive COVID-19 test and died within (equal to or less than) 28 days of the first positive specimen date" resulting in a reduction in the cumulative number of COVID-19 deaths by 13% (5,377/42,072), with a marked separation from the earlier measures from late April 2020 onwards (Garrett, 2022). Richardson (Richardson, 2022) explored similar issues in her presidential address noting the "arbitrary chosen time interval".

No definition is perfect, and the new measure could easily exclude a person who had a prolonged infection. Importantly the definition states "with a laboratory-confirmed positive COVID-19 test" which does not imply a causal link. The approach to identifying those deaths with a causal link was to access registered deaths. However, there was also some early confusion here when data was aggregated and reported. In particular, the difference between date of death and date of registration of death, with the latter clearly lagging the former. The former, provided by Public Health England, excluded deaths outside of the NHS with the latter, produced by the ONS, included all certified deaths (usually within 5 days of death). There were also important differences between England and the devolved administrations – notably Scotland where deaths must be registered with 8 days having been ascertained, even when cases are referred to the coroner. In England and Wales when deaths are referred to the coroner, the fact-of-death is not registered with the ONS until the cause of death has been determined meaning deaths can be registered months later than when they occurred (UK Covid-19 Inquiry, 2023).

As the pandemic progressed, excess deaths became another important measure. Such a measure when estimated over a period of year might be thought of as a measure of the overall high-level impact of everything - including the direct and indirect effects (both positive and negative) of NPIs (including lockdowns), treatments and vaccinations. It conceivably also enables international comparisons to be made. Age adjustment is required, but the fundamental question than arises is which baseline to use, and how to measure the effect of a pandemic that spans multiple years. In other words, in excess compared to what? The point here is that if the definition is clear and transparent at the outset then the limitations can be explored, and the data interpreted and used correctly – avoiding confusion due to meta data only revealing itself later. Addressing such concerns, the ONS has now developed standard methodology that adjusts for population growth, ageing and reflects recent trends in mortality rates that has the potential to become an international standard (ONS, 2024).

Despite these well documented issues, the tried and trusted data generating methods of random sampling and randomisation brought a rigour that changed the course of the pandemic – providing evidence to support policy.

Sylvia Richardson (Richardson, 2022) referenced the use of random sampling during the pandemic "by which the link between symptoms and testing can be broken". Indeed, the UK was the world leader here with two infection prevalence surveys conducted in parallel. Both used random sampling but with important differences.

The REACT-1 survey (Elliott et al, 2023) used a random sample design. It was limited to England with the aim of estimating infection prevalence over time, by person and place. The sampling frame was named individuals agreed 5 years and older from the National Health Service list of general practitioners – regarded as "near-universal coverage". Data were collected monthly, over a period of 2-3 weeks, from 1st May 2020 to 31st March 2022. Those individuals selected were invited to register online or by telephone and those that registered received a questionnaire and a swab kit by mail. Each round was constituted a new random sample (cross-sectional design). Over the course of the study, over 2.5 million swabs were obtained from just over 14 million invitations with an overall response rate of 17.9%.

The Office for National Statistics (ONS) Covid-19 infection survey used a wider UK sampling frame, but it was also narrower than REACT-1 since it limited the sampling frame to UK households. Importantly is collected some longitudinal data (with repeat samples from the same households through time) and recorded symptoms and infections for all those in the household. The latter generating data to investigate transmission within a household which REACT-1 could not.

Both surveys brought regular estimates of disease prevalence that began to reliably document the course of the pandemic in the UK, influence policy and inform public debate. Notwithstanding the success of these two surveys, low and deteriorating response rates were a problem in both cases. Despite scientific rigour resulting from the properties of random sampling, potential bias resulting from non-response remained a challenge – a point revisited in section 3.3.

Randomisation was successfully deployed to investigate the use of re-purposed drugs in those hospitalised due to infection and to develop vaccines – resulting in evidence from causal models that turned the tide of the pandemic.

The UK's RECOVERY trial (UK Research and Innovation, 2024) is one of the most impactful clinical trials of recent years, initially re-purposing drugs to treat COVID-19 and then evaluating them in an experimental setting. Led by the University of Oxford, it was set up in March 2020 within 6 weeks of funding. It was a randomised, parallel group study that used a platform design (an adaptive clinical trial) with a master protocol that had multiple treatments and it was able to add promising treatments, including newly developed drugs, or drop ineffective ones through time – although it was not double-blind. Its breakthrough in June 2020 was to demonstrate the effectiveness of dexamethasone in reducing 28-day mortality and later in the effectiveness of tocilizumab and the experimental monoclonal antibody combination casirivimab/imdevimab. Of equal importance in June 2020, it confirmed that the anti-malarial, Hydroxychloroquine (HCQ), was ineffective. Observational data from China, early in the pandemic, pointed to its effectiveness (with additional data from countries in Europe and the US) but the findings were subject to much debate since the data had not been generated from randomised trials. Interestingly, an observational study published in the Lancet concluded that HCQ was ineffective but was later retracted leading to the Lancet to review it processes with more attention given to data provenance and proof of data sharing

agreements. Notwithstanding the findings of RECOVERY, it certainly strengthened the case for using an experimental framework to generate evidence to support policy (Garrett, 2022).

Vaccine development was also a major success story from the pandemic with positive clinical trial results emerging and UK emergency use authorisation granted on 2nd December 2020 for the first vaccine within an unprecedented 12-month timeframe. The vaccine demonstrated unexpectedly high efficacy and cleared the well-defined regulatory criteria with ease (relative effect ≥50% with corresponding lower confidence limit >30%). Other vaccines with positive outcomes followed swiftly. The phase III clinical trials for the vaccine were large but also simple parallel group designs that randomised volunteers to vaccine (typically two doses separated by two weeks) or control (in most cases placebo) in a double-blind manner with follow-up to record occurrences of symptomatic infection. Following the regulatory guidance, the study designs across a range of vaccines were very similar and enrolled in the region of 30,000 to 44,000 subjects per study, providing safety data orders of magnitude larger than would typically be found in drug development for a treatment. The trials demonstrated that the vaccines provided protection against symptomatic infection for those vaccinated and were safe in terms of risk benefit. However, they did not provide evidence on onward transmission of SARS-CoV-2 by those in the clinical trials to those not in the clinical trials. Jump to 11th October 2022, and this proved controversial.

A European Parliament Dutch MEP's intervention illustrates how a lack of understanding of the data generating process can easily lead to false claims. In a post viewed more than 13 million times, Rob Roos stated how a pharmaceutical executive revealed that Covid-19 vaccines were not tested for their impact on transmission prior to release. "This is scandalous. Millions of people worldwide felt forced to get vaccinated because of the myth that you do it for others. Now this turned out to be a cheap lie." In a subsequent email (Full Fact, 2022) he wrote: "Governments worldwide have introduced Covid mandates and passports that had an enormous impact on millions of people. They did so by explicitly arguing that vaccinated people cause less transmission of the virus" and "…proves this was an assumption by governments for which no evidence had been provided".

Notwithstanding that vaccine developers were not required by the regulators to test the impact of a vaccine on transmission prior to its release, it is perhaps not unreasonable for the public to ask why the vaccine trials did not look at transmission. Randomised trials typically assign individuals to interventions and protocol endpoints (in the case of vaccine development, the occurrence of symptomatic infection) are recorded post-randomisation for those individuals. Importantly those individuals consent to take part in the clinical trial, receive an intervention (the vaccine or a placebo) at random and have endpoints measured. However, these endpoints are not recorded for all the people who are not in the trial, who happen to interact with the individuals in the trial. It would be both impractical and unethical to do so. However, it follows indirectly that if a vaccine reduces infection, then it should also reduce transmission, since an uninfected individual cannot infect someone else, but importantly the direct effect cannot be quantified from these clinical trial designs. It is also conceivable that behavioural change could lead to the increased circulation of asymptomatic individuals (Senn, 2022).

In fact, evidence for reduction in transmission due to vaccination did subsequently come through observational studies using record linkage – notably the Scottish Health Care Workers (HCW) household transmission study between December 2020 and March 2021 of 300,000 people (Shah *et al*, 2021). In this study the rate of infection for people that lived with HCWs was at least 30% lower when the worker had been vaccinated (mostly with a

single dose). It was noted that this was likely an underestimate as households were exposed to others beyond the HCWs.

It is unclear if the MEP was being deliberately devilish - trying to mislead or simply not investing the time to understand the regulatory requirements and the clinical trials designs. Perhaps the pharmaceutical representative did not explain well or communicate effectively. I suspect all three.

3.3 Official statistics

The Society has a long-standing history of being actively involved in discussions around inflation including contributing evidence to the House of Lords Economic Affairs Committee on the "The Use of the Retail Prices Index" in 2018. At this time, it was noted that "no one consumer price index can meet all user needs" and that a measure for macroeconomic purposes, such as targeting inflation, was different to one used to understand the impact on households. Indeed, the Society has campaigned for greater granularity in relation to household inflation in terms of how different groups in society are impacted by rising costs (Astin & Leyland, 2015). The campaign focussed on the importance of the production and regular publication of household cost indices (HCI) that breakdown household cost inflation by income groups (deciles), housing status (e.g. private renters, outright owner occupiers), whether the household has children and retirement status. (ONS Statistical Bulletin, 2024). At the time of writing, the ONS had published its second set of quarterly HCIs which has been welcomed by the Society.

At its heart is the question as to whom a measure of inflation applies and the resulting consequences for various groups in society if a single measure is uniformly applied without understanding the impact – particularly on the most vulnerable groups in society. HCIs are intended to measure inflation as experienced by households - that is, "how much a household's disposable income would need to rise to compensate for inflation" (Astin & Leyland, 2023). They are intended to complement, not replace, macro measures of inflation that comply with international standards. A key element of the Society's recent focus has been to point out that the Consumer Prices Index (CPI) does not include mortgage and loan interest payments, and during a time of rising interest rates, the impact on households is likely to be underestimated and be more variable leading to a potential disconnect between CPI and what specific groups were experiencing. Amongst other points, CPI also gives more weight to the expenditure patterns of wealthier households. In this respect, the detail is important and an awareness of the consequences - that is, how is inflation defined and measured, and given the detail, how should different measures, that go deeper and are more granular, be used to inform debate and influence policy making beyond CPI. The February 2024 quarterly release of the 2023 data illustrate the impact and show how during the recent period of high inflation those households with children and those with mortgages fared much worst. In December 2023 the annual rate was 5.5% for households with children compared with 4.8% for those without. The highest inflation rate was 6.3% for mortgagor owner occupier households compared to 4.0% outright owner occupiers. Overall, the all-household HCI annual rate was higher at 5.0% compared with CPI at 4.0%.

Such thinking in relation to granularity is not new and Harold Wilson in his presidential address (Wilson, 1973) referred to developments including a "new Index based on the expenditure pattern of pensioners" and later "developments based on social surveys, and the new General Household Survey in particular, will enable us to form reliable estimates about the impact of individual events within the family". Claus Moser (Moser, 1980)

expressed a different view however in his presidential address "… I myself resisted proposals to introduce price indices for special groups or regions partly on the grounds that they might undermine confidence in the RPI itself. Experience in other countries has shown how a battery of indices, or even two, can be used competitively".

Inflation is one statistic, but there are many other statistics where the public awareness and understanding of the statistics is potentially lacking. UK based research through interviews, focus groups and surveys (Runge, 2023) points to the public's "deep distrust of economic statistics" and their struggle to reconcile those statistics with "the hardship they see in their local community". In contrast a survey (NatCen, 2024) found that in 2023, 90% of respondents agreed that the statistics produced by the ONS were important to understand the country and 87% trusted the ONS – a percentage that was markedly higher than for Government (31%) or the British media (25%). However public misconceptions and false confidence prevail in relation to understanding economic measures. Runge found that the public views unemployment as binary (working or not working) whereas the official measure splits non-working into those actively (unemployed) versus not-actively (economically inactive) seeking work. In a recent survey "two-thirds of respondents of the UK public described their understanding of the unemployment rate as good, higher than for GDP, inflation and interest rates". Runge noted that at the time, the mismatch was 20 percentage points between the UK's official unemployment measure (4%) and the public's binary measure (24%). He proposed that labour statistics should be explained better – in particular, the non-working split with more detail on those not actively seeking work (due to long-term sickness, caring responsibilities, early retirement, or further study). He also noted that in surveys "participants often question why people who are working very few hours are included in the employment figures, and more broadly the notion that all jobs count equally, regardless of job security and pay".

Of course, there is a need for statistics to comply with international standards at the macro-economic level (that support international comparisons, macro-economic management, and planning, and comply with economic theoretical frameworks). However, as Runge noted, these macro-economic measures created for economists, policy makers and researchers "were not designed for public communication". As such, there remains a need to supplement these important macro-economic measures with statistics that are better understood and more easily communicated – ones that are more intuitive and useful to the wider public. This is a point that I will return to later.

There is also the question of where do the data come from – that is, how are the data generated? Runge found that the public "assume that the headline employment figures are based on claims data and tax data, which come from government departments rather than large surveys from an independent statistics agency". Indeed, other research (ADR UK, 2020) has demonstrated that the public generally assume that data collected by government for a variety of purposes (that is administrative data) are used, shared, and linked for the public good – and are somewhat surprised when they are not.

In fact, employment statistics are generated from the Labour Force Survey (ONS, 2024) which is the largest household survey in the UK. Data is collected over the phone or in some cases face-to-face and the questionnaire includes a range of topics beyond employment status and including education, training and health. It is non-compulsory and uses the Royal Mail's Postcode Address File with households selected at random. Despite the data generating process being based on sound statistical principles, employment statistics have recently lost National Statistics status (being re-badged as official statistics in development) due to smaller samples and resulting discontinuity. This led the ONS to state

that “estimates of quarterly change should be treated with additional caution” and further that they should be used as part of a “suite of labour market indicators” including administrative data. Such challenges are not new and in Harold Wilson’s presidential address (Wilson, 1973), he warns of “vacancy figures, which are now treated by some media commentators as representing the word of God”. He continues “My conclusions, which are set out in Beveridge’s *Full Employment in a Free Society*, were that the total figures were meaningless, as were variations, behaving in an exaggerated way when the labour situation was locally tight: nevertheless trends in them were – and are – significant over a few months”.

Governments and their statistical agencies are undoubtedly in a transition period where the traditional data generating methods based on sound statistical principles such as random sampling are being undermined by low survey response rates (as discussed in Section 3.2) whilst increasing amounts of administrative data is becoming available. To transition to new methods and different data in a digital world requires short-term duplication of effort to maintain the continuity of important macro-economic measures over time. It also points to an opportunity to work with users to provide more granular measures (broader and deeper) that meet with wider public needs. Although there have been announcements of government investments in AI research in the UK (circa £1 billion announced in the 2023 UK budget), these investments should not be at the expense of investment in the core data infrastructure of UK and the bodies that produce important information. So let us now turn to AI.

3.4 Artificial intelligence

Digitalisation marked an important epoch in data generation. The conversion of text, images, sound, etc. to data has transformed the field of data – broadening the data corpus in unimaginable ways. The data generating processes are so ubiquitous, and frequently automated and passive, that they are all too often hidden. New terms have been created such as data ingestion, data wrangling and data lake in the world of data science, and Artificial Intelligence (AI) has become the buzzword. As John Pullinger (Pullinger, 2013) noted in his presidential address “Our wonderful information age has brought with it information overload” and “The data deluge is undifferentiated”. He asks: “How do we judge statistical provenance?” Fewer than 25 years ago, an article in the Daily Mail (Chapman, 2000) stated the “Internet ‘may be a just a passing fad as millions give up in it’”. It quoted researchers stating that “e-mail, far from replacing other forms of communication, is adding to an overload of information”. Of course, it now dominates communication and adds to the overload. It seems a world away from the experiences of Harold Wilson (Wilson, 1973) when he described his “days with Beveridge, when I had to do two hours of forced statistical labour before breakfast” estimating amongst other things “seasonal variations in the monthly unemployment figures by industries (1927-37)” with a “slide-rule contraption constructed of slats of thin wood, equivalent to a flat rule 37 feet long: I used a Otis King, equivalent to 66 inches”. In fact, Wilson noted that he still used it in 1973 whilst acknowledging that “Equipment today is, I understand it, more sophisticated.”

Through the lens of AI, data can be used to inform, recommend, persuade, mislead, decide, or simply entertain – and the risk associated depend on the purpose and the impact. There are concerns expressed around existential threat, the need to regulate (or not) and a plethora of reports, guidelines, and white papers that in themselves perhaps reflect information overload. In one white paper (DSIT, 2023) “proportionate” was used 68 times, disproportionately so, one could argue - but the point is well made. AI is such a wide topic

that a proportionate response is required, and statisticians and data professionals are well placed to take that measured and proportionate view and to assess risk.

In his presidential address, Peter Diggle (2015) noted that "The rise of data science could be seen as a potential threat to the long-term status of the statistics discipline" arguing "that, although there is a threat, there is also a much greater opportunity to re-emphasize the universal relevance of statistical method to the interpretation of data". It is stance that I very much agree with as debate has progressed from data science to AI. Statistics is as central to AI as it is to data science – producing outputs from inputs via statistical algorithms. It follows that as part of this process, statisticians and data professionals have a key role to play in focussing attention not only on the outputs from AI, but also the inputs – on the data, and the data about the data. My sense is that the inputs have proportionately received far less attention than outputs of AI. It is one area that statisticians understand incredibly well – and it is one that raises issues around the data generating process, around representativeness and diversity and also around ethics, consent, privacy, copyright, and what constitutes fair access.

Generative AI (Criddle, 2023), with its Large Language Models (LLMs) has generated much discussion during my presidency – with the launch of ChatGPT in November 2022. Essentially next word prediction tools, these LLMs use deep learning and neural networks whereby vast digitalised inputs become outputs that then become inputs in a multilayered approach that estimates a mind-blowing number of parameters at odds with the traditional statisticians' principle of parsimony. Working through the layers, the LLMs produce the final outputs – extending from text to producing pictures and sound that humans can consume.

Ultimately the final LLM outputs have the potential to become digital inputs again as they become part of the corpus of data to train new LLMs. This occurs if the outputs are published papers, or reports, or pictures or sounds that are stored and accessible. Which brings into play the topic of AI hallucinations – whereby LLMs quote references and other material that simply do not exist – that are made up or false. In some respects, this may be regarded as mis-information – information that is false but not deliberately so. This is different to dis-information that is deliberately created to deceive. Outputs can be created from inputs for entertainment or amusement purposes – writing a poem in the style of Robert Frost or creating a picture of yourself in the style of Banksey – but this is very different from fake news and fake images that may be created to deceive or embarrass. This is truly the work of the devil in disguise.

Hallucinogenic AI has the potential therefore to undermine knowledge – to degrade it over time at an increasing rate as LLMs are increasingly adopted. This may represent the greater existential threat - the slow erosion of knowledge. The provenance of data becomes an increasingly important and complex topic as a result – the data about the data in a fully digitalised world. A similar point has been made by David Hand (Hand, 2018) – "that the ability to follow-up the source, and the possibility that one might, should always be there". The traceability of the inputs to the outputs also becomes increasingly important where the outputs impact others, particularly if the impact is in relation to a decision and there is limited human oversight. This could include a medical decision based on the reading of an MRI scan, a decision on who is eligible for housing benefits, or the treatment of an individual in the judicial system (Tarran, 2023)

Provenance is also important in relation to consent, privacy, and copyright and lawsuits are arising in relation to what constitutes fair access. It also extends beyond copyright to plagiarism if content is produced from LLMs that is essentially word-for-word identical to news articles published elsewhere behind paywalls. Referred to as "memorizing content" by

the New York Times (chatgptiseatingtheworld.com, 2023), text of greater than 100 words is reproduced in identical order by ChatGPT-4 to their own news story but for one word change. Separately this raises questions around how LLMs are working – and whether they are taking short-cuts to answer some of the questions posed – and if so, where to?

The Society is creating an AI Task Force to plan and link its work aligned with the Society's 2024-2029 strategy. It will identify issues and opportunities where the Society might impact public discourse and decision making around AI and to lead the Society's policy work whilst strengthening relationships with adjacent organisations to increase impact (Royal Statistical Society, 2024). It will also bring in the practitioners' voice - since the Society has access to engaged and well-informed fellows who can bring unique insight and should enable the Society to have a stronger voice in this area. As John Pullinger (Pullinger, 2013) stated: "The professional statistician offers users of statistics confidence in the provenance of and conclusions from the data".

It is now time to look beyond the examples to consider the Society's relationship with data and those that work with data, to consider matters of trust and how the Society exerts influence by working with others.

4. The Society and data

As researched thoroughly by John Pullinger (Pullinger, 2013) for his Presidential address, the Society's relationship with data goes back to its origins and its founding fathers – including the wheatsheaf symbol "to show that their passion was to gather data, so that they could be threshed to reveal the golden corn, the value hidden in the fields that lay all around them". William Beveridge, Charles Booth, Austin Bradford Hill, Thomas Malthus, Florence Nightingale, are all fellows that we associate with data. There are simply too many to reference.

In 2013, the five-year strategic review introduced the Society's strapline of "data, evidence, decisions". If I recall correctly, the word "data" was a suggestion from Rita Gardner, then Director, Royal Geographical Society (now Chief Executive, Academy of Social Sciences) and an external member of the Society's Long-Term Strategy Group that I chaired. The aim was to capture the mood with the increased interest in data and data science. It certainly resonated with the group and alongside evidence (for gravitas) and decisions (for impact) it became our strapline.

Subsequently the Data Science Section (now the Data Science and AI Section) was set-up in 2017. The paper I presented to Council was directed to switching the position whereby each section was to incorporate data science into their own section to instead one where data science had its own focal point – including to attract data scientists from academia, government and importantly practitioners from industry. It was at the time that John Pullinger was setting up the ONS's Data Science Campus in Newport and there was an active community of data scientists arranging "beer and pizza" meetups in London and elsewhere. Following on the Data Ethics and Governance Section was formed in 2020 after three years as a special interest group. Sylvia Richardson convened the Society's Data Science Task Force to investigate "how the RSS can deepen and extend its impact on the field of data science and its offering to fellows who work as data scientists". It was focussed on delivering three main initiatives – establishing professional standards for data scientists through the Alliance for Data Science Professionals, a new online data science journal and an online data science platform (Real World Data Science).

One question that often arises, is how far does the Society move towards data – and does that imply a move away from statistics? As Deborah Ashby in her presidential address reflected on our 1887 charter (Ashby, 2019) "from its earliest days, the Society has been about using data for the public good" and is some ways the challenge is how to ensure that the Society continues to represent methodologists who may not see them themselves as working with data. We want to support and nurture methodological research and bring career young researchers into the Society.

This tension has been carefully managed in the Society's 2024-2029 strategy (Royal Statistical Society, 2024) through extensive consultation and iteration by the long-term strategy group. chaired by Jennifer Visser-Rogers, and it finally arrived at the collective term: "Statisticians and other data professionals", with an explainer for the term data professionals. That is, "The term data professional is being increasingly used as an umbrella term to describe statisticians, data analysts, data scientists, data architects, and data engineers. Data professionals work across a variety of sectors including academia, government, civil society, business and industry, and the media, and our membership reflects this. We have taken an inclusive approach in using the term 'data professionals'".

What remains clear, is that the Society continues to be a broad based and welcoming society that is keen to represent and support a range of interests from academia to government to industry.

5. The Society and trust

David Spiegelhalter's presidential address (Spiegelhalter, 2017) spoke to Trust in Numbers – bringing together issues related to reproducibility and living in a 'post-truth' society. He noted the "association with claims of a decrease in trust in expertise" and "the use of numbers and scientific evidence".

As has been shown earlier, the detail is important to enable others to reproduce results and to determine whether claims are valid. Digging into the detail behind the numbers can reveal previously unknown limitations – or known limitations not communicated well. David argued that statisticians and the Society have an "essential role both in improving the trustworthiness of statistical evidence as it flows through the pipeline, and in improving the ability of audiences to assess that trustworthiness". It is a position that is hard to disagree with as illustrated in the earlier sub-sections. David speaks to critiquing in terms "of the number itself", "of the conclusions drawn" and "of the source and what we are being told". David describes three ways to increase the trustworthiness of statistical evidence - "change the communication structure, improve the filters for the information being passed and improve the ability of audiences to check trustworthiness".

Perhaps another aspect of trust is the confidence that the public has in being provided with the information in a timely manner to hold government and other bodies to account – and being provided with the information that is important to them, including more granular information. This takes us into the realms of user requirements. A past president of the Society, Denise Lievesley, has recently completed her Independent Review of the UK Statistics Authority (Lievesley, 2024). Notwithstanding the thoroughness and effort required to produce such a review, it is worth highlighting a select number of recommendations that feed into the themes of this address. Denise's first recommendation (Recommendation 1) is the establishment and delivery of a Triennial Statistical Assembly that "would involve key organisations inside and outside Government and across the four Nations". This would determine the needs for statistics including "the private sector, government departments,

local government, academia, think tanks and media representatives". Importantly it is aimed at "identifying data gaps and ensuring users can hold the statistical system to account on the delivery of the programme of work". The Society, that has long supported the need for users to have a voice and has welcomed this recommendation. It is actively and constructively engaged in the process as a result. Back in 1980, Claus Moser stated: "I felt the need for a user forum which would help government statisticians develop their work for government as well as for the community as a whole, and which would constructively increase public awareness of the benefits of statistics". Various forums have indeed been created over the years, and it is now time to create one that meets the needs of the digital world that facilitates going broader and deeper.

Denise also recommends (Recommendation 5) creating common standards and improving harmonisation across the four nations "to strengthen the Concordat of Statistics". Such a move inevitably makes communication simpler and enables consistent UK-wide statistics to be produced (as illustrated earlier in section 3.2). Recommendation 9 includes actions around communication including ensuring "a better understanding of the levels of uncertainty around official statistics, particularly economic to reduce public (and government) surprises to revisions". Furthermore, she recommends "building partnerships with organisations that foster relevant communication expertise to improve engagement with the wider needs of users". These are all excellent recommendations that have the potential to increase the trustworthiness of government statistics. As noted by Runge: "There is a role for statistics agencies, statisticians and economists to take more ownership of the statistics and become more visible to the public, including reinforcing the message that the numbers are impartial, independent and transparent".

Holding others to account and the need to address trustworthiness goes beyond government and the same principles should apply to medicine, climate change and AI. They apply to pharmaceutical companies, research bodies and "Large Tech" given their influence on our day-to-day lives. For clinical research, the Sense about Science ALLTRIALS initiative (launched in 2013) called for "all past and present trials to be registered and their full methods and summary results reported" and "in March 2023 the MHRA announced the introduction of a new legal mandate that calls for all clinical trials in the UK to be registered in a World Health Organisation (WHO) public register" (Sense about Science, 2024). It strikes me that world of Large Tech and AI is the one that is mostly lacking.

6. The Society and collaboration

The Society's 5-year strategy (Royal Statistical Society, 2024) introduces, for the first time, a set of values to guide the Society's work and our wider membership, providing purpose and direction. These four values are inclusive, collaborative, impactful and progressive. Although Collaborative is simply stated as: "we work in partnership with other organisations and individuals" it is an enabler of the other three, in that we can be more inclusive, impactful, and progressive by working with others.

Recent examples of that collaboration include working with Citizens Advice on HCIs and the Council for the Mathematical Sciences to form the Academy for the Mathematical Sciences. The Alliance for Data Science Professionals is an example where the Society has worked collegially with seven other societies and institutions to define "the standards need to ensure an ethical and well governed approach so the public, organisations and governments can have confidence in how their data is used". These standards refer to individuals, accreditation of university courses and certification for individuals and education providers.

Through the Society's royal charter, the aim is to be able to offer Chartered Data Scientist through all eight members of the Alliance by the end of 2024.

The ubiquity of data and statistics means that the Society has an opportunity to contribute widely – perhaps more widely than ever before. It is an opportunity but also a challenge since resource and time is not unlimited and we are highly dependent on our fabulous members who volunteer their time and expertise. As noted by Sylvia Richardson in relation to the pandemic, "Explaining collective risk would be best tackled from multidisciplinary collaborative efforts including social and environmental scientists as well as statisticians and Health experts". What is true for the pandemic is true for numerous other areas and it is wonderful to observe the contributions of fellows in so many different areas – leading to impact through their interactions with others.

Before providing some closing remarks, I promised in the introduction to speak to connections – to some of the earlier presidents' addresses that I sought out.

7. Connections

I have always been intrigued that Harold Wilson, the British Prime Minister (serving 1964-1970 and again 1974-76) was also a president of the society (1972-73). Lord Wilson was member of parliament for the Huyton constituency in Liverpool for 33 years until 1983, a period that coincided with my birth and schooling. A statue of him sits in what is known locally as Huyton village. Lord Wilson's presidential address was published in 1973, 50 years ago (as I started to write) and is a useful reference point to discover how things have changed over that period.

Sir Claus Moser, a previous Director of the Central Statistical office (forerunner to the Office of National Statistics) was President between 1978-80 and his address speaks to being interned as a German Jewish refugee at a camp in Huyton in 1940 (Moser, 1980). Claus Moser's address was in 1980, the year I went to university and studied Statistics for the first time as part of an Economics degree.

So, what is the Lord Stanley connection? The Stanley family owns the Knowsley estate and Huyton is in the borough of Knowsley.

8. Concluding remarks

Digitalisation has created a world awash with data, from sources almost unimaginable fifty years ago. Alongside computing brute force, digitalisation enables society to go broader and deeper than ever before. It has the potential for increased transparency and to empower.

Correspondingly, statisticians and data professionals have an increasingly important role to play as advocates for data generation based on sound statistical principles since how the data is generated has fundamental implications. It requires statisticians and data professionals to remember to "get out more" – to understand why and how the data is being generated – particularly in the age of passive data generation with its associated traps. They also need to understand what is not being collected and why. The risk of false precision due to sheer data volume means that bias becomes more important than ever before. (It may be considered remiss not to have covered uncertainty in this address. This is for no other reason than the focus is on data rather than analysis and because precise measurements of the wrong thing is meaningless.) It requires statisticians and data

professionals to think about design – to influence the data generating process and to implement data standards. It entails communicating well and documenting with rigour.

The Society has a role to play in helping the public to navigate the increasingly complex world of data and as part of the Society's 2024-2029 strategy its goal for "Supporting public understanding and engagement" states that "People have an understanding of the data and statistics that influence their daily life decisions, their work and the world around them, and feel empowered to meaningfully engage with issues". This is alongside the goal of "Championing the public interest" stating "Societal decisions are informed and improved by the appropriate use of data and statistics that are reliable, that are used responsibly, and that are relevant to society's most important questions". As the pandemic has shown, the public are interested in the data, but they need access to trusted sources and trusted voices.

Claus Moser stated: "Moreover, the public of the 1980's will be better educated and will more consistently challenge the decisions of government. They will expect to monitor government success and will expect readily accessible, convenient and intelligible statistics with guidance on quality of data and on meaning." Over four decades on, the public are even better educated, and it is not simply governments that they want to challenge but other public bodies and corporations. Runge found that "people are hugely interested when presented with engaging data" but "also frustrated that don't come across this type of information on a regular basis" (Runge, 2023). It is all too easy to think that the best approach is to limit what is made publicly available – to brush-over the detail, to avoid the nuances. I prefer to think that the world has moved on, and that there is an increasing appetite for a more open approach. There is a risk of cherry-picking, but that should not be an excuse for limiting the availability of useful information. Trusted sources are required, and the Society needs to be seen as one of those sources - a champion for the public understanding of data and statistics.

As digitalisation enables AI, data provenance will become increasingly important - the traceability of data on their journey from input to output – and to input again, potentially as part of a never-ending cycle. In this respect it mirrors the fact that data becomes statistics that become data for others to use. In my view "inputs" is an area that has not received enough attention and one that statisticians understand incredibly well. One concern with AI is that traceability becomes impossibly challenging to undertaken due to the sheer volume and multi-layering, while hallucinogenic AI has the potential to slowly erode knowledge and undermine aspects of society over time – representing a greater long term existential threat than the oft-quoted rogue robots. The Society cannot solve these issues - but it can strengthen its voice by working with others to raise important considerations and exert influence. Topics such as representativeness, fairness, consent, and privacy are all parts of our natural purview. It is timely that the AI Task Force is being formed as part of our five-year strategy to bring together the unique expertise that we undoubtedly have across industry, government, and academia available through our broad Society membership.

The Society has an important constructive role to play to support and represent the users' voices. The triennial review proposed by Denise Lievesley in her Independent Review of the UK Statistics Authority is an opportunity for the Society to have constructive engagement with the UKSA around government data and statistics. It is also an opportunity to support calls for greater investment in the UK's data infrastructure such that a well-resourced ONS and Government Statistical Service can continue its transition from surveys with low response rates to a mix that incorporates fit-for-purpose administrative data whilst also meeting society's growing needs. It is important that UK government's bold investment in AI does not come at the expenses of the provision of core data services that would simply be

false economy. However, there is perhaps greater opportunity for researchers and others to pull government data rather than simply having ONS and other government bodies push it. With controlled access to clearly defined, well documented data and a co-ordinated effort by accredited researchers and others, more unmet users' requirements may be satisfied. More generally our aim should be to be viewed as the users' champion regarding using data to hold other bodies to account.

Finally, as Moser (Moser, 1980) stated: "The methods, concepts, standards, definitions and all other aspects of statistics, surveys and censuses are the statistician's professional province." Expanding this to include other data professionals, it is time for the Society to re-enforce that message. The Society embracing data and data professionals does not represent a shift away from statistics, rather it reaffirms where it all started - it re-establishes the link to our roots. This comes at a time when we look to leave Errol Street, our home for approaching 30 years. The Errol Street office was originally built as a school in 1890 and it is unclear where the playground would have been, but regardless the world is one big playground for statisticians, data professionals and the Society. It is hard to imagine a more exciting time for the Society than to be part of a world awash with data – even if a few battles with the devil lie ahead.

Acknowledgements

I am indebted to David Hand, Stephen Senn and David Stephenson for providing valuable and thoughtful comments on a draft version of the address. All errors and omissions are my own.

References

ADR Administrative Data Research UK (2020). Trust, Security and Public Interest: Striking the Balance. A review of previous literature on public attitudes towards the sharing and linking of administrative data for research Available at: https://www.adruk.org/fileadmin/uploads/adruk/Trust_Security_and_Public_Interest-_Striking_the_Balance-_ADR_UK_2020.pdf

ADR Administrative Data Research UK (2024). What is administrative data? Available at: https://www.adruk.org/our-mission/administrative-data/

Arrhenius, S. (1896) On the Influence of Carbonic Acid in the Air upon the Temperature of the Ground. *Philosophical Magazine and Journal of Science*, 41(5), 237-276.

Ashby, D. (2019) Pigeonholes and mustard seeds: growing capacity to use data for society. *Journal of the Royal Statistical Society: Series A (Statistics in Society),* 182(4), 1121-1137

Astin, J and Leyland, J. (2015) Towards a Household Inflation Index: Compiling a consumer price index with public credibility. Available at: https://rss.org.uk/RSS/media/News-and-publications/Publications/Reports%20and%20guides/Towards_a_Household_Inflation_Index_May_2015.pdf?ext=.pdf

Astin, J and Leyland, J. (2023) Measuring inflation as households see it: next steps for the household costs indices. Available at: https://rss.org.uk/RSS/media/File-library/Policy/2023/Measuring_Inflation_as_households_see_it_executive_summary_Jan_2023.pdf?ext=.pdf

Betts, R.A., Belcher, S.E., Hermanson, L., *et al*. (2023). Approaching 1.5 °C: how will we know we've reached this crucial warming mark? *Nature*. 624, 33-35.

Breslow. N.E. (2001). Statistics in the life and medical sciences. In: Raftery AE, Tanner MA, Wells MT, eds. *Statistics in the 21st Century*. Virginia: Chapman and Hall/CRC,1-3.

Chapman, J. (2000) Internet 'may be just a passing fad as millions give up on it'. Daily Mail. 5th December 2000.

Chatgptiseatingtheworld.com (2023). New York Times sues Microsoft OpenAI for alleged copyright infringement. Available at: http://chatgptiseatingtheworld.com/2023/12/27/new-york-times-sues-microsoft-openai-for-alleged-copyright-infringement-download-complaint/

C3S Copernicus Climate Change Service (2021). New decade brings reference period change for climate data. 9th February 2022. Available at: https://climate.copernicus.eu/new-decade-brings-reference-period-change-climate-data

Criddle, C. (2023) What is artificial intelligence and how does it work. Financial Times. 20 July 2023.

Department for Science, Innovation and Technology (March 2023). A pro-innovation approach to AI regulation. Available at: https://assets.publishing.service.gov.uk/media/64cb71a547915a00142a91c4/a-pro-innovation-approach-to-ai-regulation-amended-web-ready.pdf

Diggle, P.J. (2015) Statistics: a data science for the 21st century. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 178(4), 793–813.

Elliott, P., Whittaker, M., Tang, D et al. (2023) Design and implementation of a national SARS-CoV-2 monitoring program in England: REACT-1 study. *Am J Public Health*, 113(5), 545-554.

Food and Drug Administration (1997). The Food and Drug Administration Modernization Act of 1997 (FDAMA) (Pub. L. 105–115-NOV,21, 1997). Available at: https:// www.congress.gov/105/plaws/publ115/PLAW-105publ115.pdf

Full Fact (2022). Claims that Pfizer vaccine wasn't tested on preventing transmission need context. Available at: https://fullfact.org/health/coronavirus-vaccine-pfizer-transmission-test/

Garrett, A.D. (2006). The role of subgroups and sub-populations in drug development and drug regulation. PhD thesis The Open University. Available at: https://doi.org/10.21954/ou.ro.0000d565

Garrett, A. (2016) Independent review of Methodology. UK Statistics Authority. Available at: https://uksa.statisticsauthority.gov.uk/publication/independent-review-of-methodology/

Garrett, A. In Hand, D.J. (2018) Discussion: Statistical challenges of administrative data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181(3), 582-583.

Garrett, A.D. (2022) The cross-over of statistical thinking and practices: A pandemic catalyst. *Pharmaceutical Statistics*, 21, 778–789.

Hand, D.J. (2009) Modern statistics: the myth and the magic. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(2), 287-306.

Hand, D. J. (2018). Who told you that? *Significance*. 15(4), 8-9. Available at: https://doi.org/10.1111/j.1740-9713.2018.01166.x

Hand, D.J. (2018) Statistical challenges of administrative data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181(3), 555–605.

IPCC (2018). Global Warming of 1.5°C. An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty. [Masson-Delmotte, V., P. Zhai, H.-O. Pörtner, D. Roberts, J. Skea, P.R. Shukla, A. Pirani, W. Moufouma-Okia, C. Péan, R. Pidcock, S. Connors, J.B.R. Matthews, Y. Chen, X. Zhou, M.I. Gomis, E. Lonnoy, T. Maycock, M. Tignor, and T. Waterfield (eds.)]. Chapter 1 FAQ 1.1. 79-80. Cambridge University Press, Cambridge, UK and New York, NY, USA, 616 pp. https://doi.org/10.1017/9781009157940.

Intergovernmental Panel on Climate Change (2023). Sixth Assessment Report. Available at: https://www.ipcc.ch/assessment-report/ar6/

Lievesley, D. (2024) Independent Review of the UK Statistics Authority. Available at: https://www.gov.uk/government/publications/independent-review-of-the-uk-statistics-authority-uksa-2023.

Medical Research Council. (1948). Streptomycin treatment of pulmonary tuberculosis: a Medical Research Council investigation. *British Medical Journal*, 2,769-782.

Molena, F. (1912) Remarkable Weather of 1911: The Effect of the Combustion of Coal on the Climate — What Scientists Predict for the Future. *Popular Mechanics, March issue,* Chicago, US. 339-342.

Moser, C. (1980) Statistics and Public Policy. *Journal of the Royal Statistical Society: Series A (General)*, 143(1), 1-32.

NatCen. National Centre for Social Research (2024). Report: Public Confidence in Official Statistics. Available at: https://natcen.ac.uk/publications/public-confidence-official-statistics

NOAA (2024). National Centers for Environmental Information. National Oceanic and Atmospheric Administration. Global Surface Temperatures Anomalies. Background Information – FAQ. Available at: https://www.ncei.noaa.gov/access/monitoring/global-temperature-anomalies

NASA (2021). The Raw Truth on Global Temperature Records. Available at: https://science.nasa.gov/earth/climate-change/the-raw-truth-on-global-temperature-records/

NCDS National child development study (2024). About. Available at: https://ncds.info/home/about/

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*. 97(4), 558-625.

Office for National Statistics. Household Cost Indices for UK household groups: October 2023 to December 2023. 26 February 2024. Available at: https://www.ons.gov.uk/economy/inflationandpriceindices/bulletins/householdcostsindicesforukhouseholdgroups/latest

Office for National Statistics. Role of owner-occupiers' housing costs in Household Cost Indices, UK: 2023. 26 February 2024. Available at:

https://www.ons.gov.uk/peoplepopulationandcommunity/housing/articles/roleofowneroccupiershousingcostsinthehouseholdcostindicesuk/2023

Office for National Statistics. Employment in the UK: 16 April 2024. Available at: https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/bulletins/employmentintheuk/april2024

Office for National Statistics. Estimating excess deaths in the UK, methodology changes: 20 February 2024. Available at: https://ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/causesofdeath/articles/estimatingexcessdeathsintheukmethodologychanges/latest

Office for National Statistics. About the census. 19 July 2022. Available at: https://www.ons.gov.uk/census/aboutcensus/aboutthecensus

Office for National Statistics (2024). Labour Force Survey. Available at: https://ons.gov.uk/surveys/informationforhouseholdsandindividuals/householdandindividualsurveys/labourforcesurvey

Our World in Data (2024). Annual CO_2 emissions. Available at: https://ourworldindata.org/grapher/annual-co2-emissions-per-country?country=~OWID_WRL

Our Future Health (2024). *How Our Future Heath works*. Available at: https://ourfuturehealth.org.uk/our-research-mission/how-our-future-health-works/

Pullinger, J. (2013) Statistics making an impact. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 176(4), 819-840.

Richardson, S. (2022) Statistics in times of increasing uncertainty. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 185(4), 1471-1496.

Rao, J.N.K. and Fuller, W, A. (2017) Sample survey theory and methods: Past, present and, future directions. *Survey Methodology*, 43(2), 145-160.

Royal Statistical Society (2024). RSS Climate Change and Net Zero task Force Explainers. The 1.5°C global mean warming target. Available at: https://rss.org.uk/policy-campaigns/policy-groups/climate-change-net-zero-task-force/climate-change-explainers/

Royal Statistical Society (2024). Strategy 2024-2029. Available at: https://rss.org.uk/RSS/media/File-library/About/2024/RSS-STRATEGY-FOR-2024-2029.pdf

Runge, J. (2023). “This should be a wake-up call” How to improve people’s faith in economic statistics. *Significance*. 20(3), 38-39.

Senn, S. (2015). Various varying variances: The challenges of nuisance parameters to the practising biostatistician. *Statistical Methods in Medical Research*. 24(4), 403-419.

Senn, S. (2022). *Dicing with Death*. Cambridge, Cambridge University Press. 293-295.

Shah, A.S.V. et al. (2021). Effect of Vaccination on Transmission of SARS-CoV-2. *N Engl J Med*. 385(18), 1718-1720. DOI: 10.1056/NEJMc2106757

Smart Data Research UK. (2024). *The scope of what we’re doing*. Available at: https://www.ukri.org/what-we-do/browse-our-areas-of-investment-and-support/smart-data-research-uk/.

Spiegelhalter, D. (2017) Trust in numbers. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180(4), 948-965.

Stephenson, D. (2024). How close are we to 1.5 degrees global mean warming? Available at: https://stormrisk.github.io

Tarren, B. (2023a). “I was pretty clear in my mind that we were into a no-going-back situation.” *Real World Data Science,* December 15, 2023. Available at: https://realworlddatascience.net/viewpoints/interviews/posts/2023/12/15/ian-diamond.html

Tarren, B. (2023b). Statistics and Data Science are at the heart of the AI movement – we want to be a strong voice in the debate. *Real World Data Science.* Available at: https://realworlddatascience.net/viewpoints/interviews/posts/2023/10/25/evaluating-ai.html.

Thomas, J.J. (1977). *An Introduction to statistical analysis for economists*. Weidenfeld & Nicolson.

Thorne, P. (2016). Chapter 5 - Global surface temperatures. Climate Change (2nd ed.) 21-35. Editors. T. M. Letcher, Elsevier

Wilson, J.H. (1973) Statistics and Decision-Making in Government—Bradshaw Revisited. *Journal of the Royal Statistical Society: Series A (General)*, 136(1), 1-20.

Wood, S. N., (2006). *Generalized Additive Models: An Introduction with R*. Boca Raton: Chapman & Hall/CRC.

UK Biobank (2024). A powerful resource to improve public health. Available at: https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank.

UK Covid-19 Inquiry (2023). INQ000183421 – Witness Statement of Andrew Garrett, on behalf of the Royal Statistical Society, dated 21/04/2023. Available at: https://covid19.public-inquiry.uk/documents/inq000183421-witness-statement-of-andrew-garrett-on-behalf-of-the-royal-statistical-society-dated-21-04-2023/

UK Health Security Agency (2022). Technical summary UK Health Security Agency data series on deaths in people with COVID-19. Available at: https://assets.publishing.service.gov.uk/media/61fb93118fa8f53893357fc7/UKHSA-technical-summary-update-February-2022.pdf

UK Research and Innovation (2024). The RECOVERY trial. Available at: https://www.ukri.org/news-and-events/tackling-the-impact-of-covid-19/vaccines-and-treatments/recovery-trial-identifies-covid-19-treatments/

Vichi, M. and Hand, D.J. (2019) Trusted smart statistics: the challenge of extracting useable aggregate information from new data sources. *Statistical Journal of the International Association of Official Statistics*, 35, 605-613.