Institute and Faculty of Actuaries

ISSN 2397-7213

# Longevity Bulletin

## From the Institute and Faculty of Actuaries

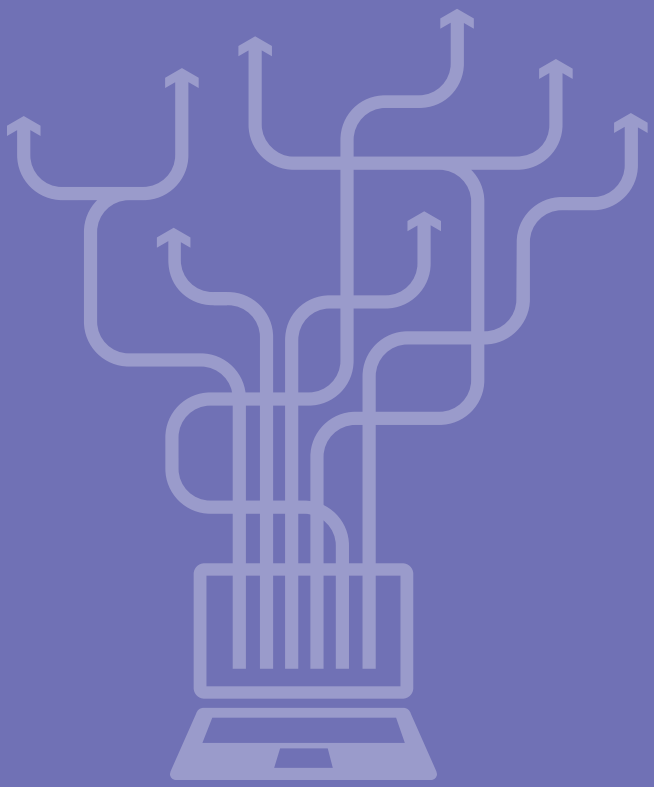
# Big Data in Health

Issue 9

December 2016

# Contents


1. Introduction by our President, Colin Wilson 3
2. Foreword by Michael Tripp 4
3. Theoretical foundations of data science, Sofia Olhede and Russell Rodrigues 5
4. Opinion: Is big data just a big hype? Peter Coveney and Roger Highfield 11
5. Routes to Diagnosis for cancer, Lucy Elliss-Brookes 13
6. Understanding longevity and morbidity risk, Elena Kulinskaya and Lisanne Gitsels 16
7. Personalised risk prediction: genomics and beyond, Philippa Brice 20
8. Big data in action: wearables, Matthew Edwards 23
9. The ethical challenges of biomedical data, Luciano Floridi 26
10. Recent developments and events 29

# 1. Introduction by our President

The Institute and Faculty of Actuaries (IFoA) is the chartered professional body for actuaries in the United Kingdom. It is dedicated to raising the profile of actuaries and the value of their skills in both established and new areas of business. The IFoA continues to embrace the value of collaboration with other professional bodies and universities to rise to the challenges and opportunities facing society.

In the past few years, big data and the potential uses of this information have been receiving a significant amount of attention. Actuaries were the original data scientists in the field of mortality and longevity. There is a growing expectation that new insights into longevity will be deeper and richer than ever. This presents a fascinating opportunity for the insurance sector and the actuarial profession. This edition of the Longevity Bulletin considers a range of examples arising from the use of big data including the development of personalised medicine, the use of wearables and novel statistical and actuarial methods for modelling mortality. You can also read about the foundations of data science and challenges arising from new forms of data.

I have great pleasure in introducing the ninth issue of the Longevity Bulletin. I would like to thank all the contributors and authors for their thought-provoking and informative articles on the topic of big data and its applications.

We hope that this issue will be read with interest by all those with technical, professional and personal interest in health, morbidity and longevity matters.

**Colin Wilson**
President, Institute and Faculty of Actuaries

**Subscribe to the Longevity Bulletin**

If you would like to subscribe to receive future editions of the Longevity Bulletin, please visit: **bit.ly/longevitybulletin**

**We welcome feedback**

Please email the Research and Knowledge team at **research@actuaries.org.uk** to make a suggestion or a comment on how we can make this publication better.

# 2. Foreword by Michael Tripp

## Chair of the IFoA's General Insurance Board and Chair of the IFoA's Modelling, Analytics and Insights from Data (MAID) Working Party

Big data is no longer the buzz word it was a few years ago, but now a fact of life with everyone involved in making sense of data on computers aware of burgeoning techniques. In its purest sense, big data means the use of data from many and varied sources – the combination of what previously might have been separate and unconnected. It also means incorporation of emerging new data sources – the obvious one being wearables, but also possible social media and other real-time measurements or interactions. One consequence is that traditional actuarial methods are being thrown up in the air and need a complete reassessment – for example in general insurance claims reserving, use of the information at full claimant level rather than the well-known grouped data triangles. Longevity investigations and research may be no different – what can today's actuary do that yesterday's could only dream about?

Looking to the future, one of the considerations is how to classify problems and thus work out what new approaches are likely to be most suitable. Peter Drucker (2016) refers to four types – the truly generic, the generic but unique to a given institution, the truly exceptional and the early manifestation of a new generic problem. Another more matter of fact – clustering/pattern recognition, predictive/causality and correlation, decision taking and optimisation leading to robotics and artificial intelligence. I'm sure readers will add their own insights.

So for longevity what are the challenges and how will tools and techniques help? The old question of 'how long will people live' perhaps morphs into how granular can we make the predictions, and reduce the uncertainty. Another one of 'what treatments produce the better (or best) results', may become how do we help improve individual diagnoses and develop optimal treatment plans for any given individual's quality of life. The list can go on – and I've not even started to discuss ethics.

In this context the research work described in this bulletin is vital.

### References

Drucker F.P. (2016). *The Peter F. Drucker Reader: Selected Articles from the Father of Modern Management Thinking.* Harvard Business Review

# 3. Theoretical foundations of data science

Professor Sofia Olhede and Russell Rodrigues, Big Data Institute and Centre for Data Science, University College London

As the volume, variety and velocity of data generation continues to increase, new possibilities for analysis are emerging in the field of data science. In this article, we discuss the theoretical concepts underpinning big data; specifically the new forms it takes, the new complex models required to describe, explain and draw inferences[1] from it, and recent advances in testing the significance[2] of apparent patterns in observations. We describe some computational and privacy constraints, and conclude by highlighting some future trends for data science.

## Introduction

In 1959, the American statistician John Tukey said, 'few of us expect to ever see a man who has analysed, or even handled, a sequence of a million numerical values…' (Tukey, 1959). Back then, the 'era of big data' in which we now find ourselves (Manyika et al., 2011) could scarcely have been imagined. Modern data analysts contend not only with larger datasets, but with complications stemming from the variety of data sources and formats available, variability in data quality and completeness, and issues related to the increasing speeds at which data are generated. Furthermore, as increasing amounts of data are human-generated, e.g. in social media and healthcare settings, privacy and security considerations become paramount (Wu et al., 2014). As in Tukey's time, the purpose of data analysis is to obtain meaningful insight, but arriving at such insight requires increasingly sophisticated approaches, which are the focus of current research in the field of data science. Herein we discuss the theory underpinning the new data analysis challenges, and highlight some key approaches being adopted to address them.

## New forms of data: the challenges

**Large and heterogeneous:** As researchers, businesses and governments seek greater insight from their data, recent decades have seen significant increases in the overall data quantities or volumes subjected to analysis. Studying very large datasets can unveil rare or unusual occurrences, e.g. uncommon genetic diseases or fraudulent financial transactions. However, large volumes alone do not guarantee useful insight: data may still suffer from contamination (alterations affecting integrity), from biased samples (non-representative of the wider population), or from missing values (due to incomplete collection or retention). Moreover, much data is collected prior to considering specific questions for investigation: in the absence of experimental design[3], the collected data may be too heterogeneous[4] – lacking the structure and consistency to shed light on specific problems. Such data are referred to as 'found' datasets, because they are often taken and fed into analyses for which they were not explicitly collected. Heterogeneous datasets, even large ones, may not equip the analyst to make confident logical inferences and conclusions, and traditional analysis methods that do not correct for the limitations in these data can produce

### Glossary

1 **Inference:** a reasoned conclusion drawn from data.

2 **Statistical significance:** the likelihood of making a given observation, assuming that the process by which the data were generated is valid. If the likelihood is very small, the observation is considered 'significant', and unlikely to have occurred by random chance.

3 **Experimental design:** a plan, formulated in advance of data collection, for the optimal collection and analysis of data, to ensure it will be possible for the analyst to infer and investigate questions of interest.

4 **Heterogeneous data:** data containing several dissimilar elements, which cannot be collectively described in a simple manner, using a single model.

inaccurate and potentially misleading deductions, underestimating the true variability of behaviour.

**High-dimensional:** As sensor technology becomes cheaper and more widely available, vast quantities of multi-modal[5] data are now routinely collectable. This can be of benefit, for instance in monitoring and predicting maintenance requirements for industrial appliances and building infrastructure, but as more quantities are tracked simultaneously over time, collected data become more highly-dimensional[6], and traditional statistical analyses are often ill-suited to identify trends in such data.

**Varying modes and high velocity:** Furthermore, data increasingly come in new, varying formats, including images, sound and video, which do not fit conventional tabular databases. Classical analyses generally study single input types, but rich opportunities exist to train computers to identify features of these media and make connections between them, for instance, predicting images based on sounds. Moreover, data are increasingly collected in real-time streams, posing challenges for storage, processing and analysis: conventionally such data would be divided into discrete batches for analysis, but new capabilities are arising for rapid online analysis.

**Privacy and security:** Finally, significant insight can be obtained by piecing together multiple datasets, and developments in data linkage enable mapping of units or individuals across different sets (Pell et al., 2014). This confers technical challenges, in maintaining accuracy, but also ethical[7] ones. Added value may exist precisely in combining sources, but the potential for de-anonymisation of personal data remains a pressing concern. Moreover, access to much data held by governments, industries and other agencies is restricted, even though it might be leveraged for public benefit. Recent years have seen significant US and UK efforts through the creation of national Open Data[8] initiatives to encourage data sharing where feasible, and in the future, new protocols for partial or time-restricted access to such data are likely to emerge.

## New types of model

To mathematically describe relationships between variables and particular observations, scientists develop models; data science is no exception. In physics, Newton's theory of gravity describes the relationships between force, acceleration and mass: one can determine how variations in any of these influence the others. This model also has predictive power: the future location of bodies can be calculated from their current location, velocity and acceleration.

Similarly, data science models provide a descriptive framework for understanding relationships between observed events, and can have predictive value. Given the new complexities described above, models are increasingly developed with contributions from both statistics and computer science, particularly the field of machine learning[9], which implants capacity for models to sharpen and make more accurate predictions for future observations. This is achieved by developing sets of complex computerised instructions, or algorithms, specially designed and adapted to process data in particular ways, and to learn from it. We shall now discuss the statistical concept of sparsity[10], and outline some recent work combining it with machine learning, for model development.

To deal with heterogeneous, high-dimensional data, the notion of sparsity has recently emerged. With very large, multivariate[11] datasets, one will more likely identify apparent relationships between particular variables where none actually exists.

To avoid this, sparsity embeds models with the expectation that most potential relationships will not actually be present (Friedman et al., 2002) (Efron et al., 2004). The most popular sparse models employ the 'lasso' approach, which penalises each included variable (Friedman et al., 2001), thus aiming to eliminate all but the simplest explanations for phenomena. As such, sparse models are a mathematical application of Occam's razor[12].

5 **Multi-modal data**: data of different forms and types, for instance analogue signals, digitised signals, images, audio, video, and data collected in various medical and environmental monitoring settings.

6 **High-dimensional data:** data having a large number of parameters associated with each individual or subject, requiring intensive computer resources to store and process. These data may also be incomplete or missing information in places, which can hinder inference and interpretation.

7 **(Data) ethics:** principles governing the collection and use of data, considering the rights of individuals, organisations and societies. Key ethical issues in data science include personal privacy, individual consent, data ownership and transparency (Royal Statistical Society, 2016).

8 **Open data:** data that is readily available to anyone who wishes to access, use or share it. The UK Open Data Institute (**www.theodi.org**) and the US data.gov initiative, amongst others, increasingly call for public data to be made widely available for social good.

9 **Machine learning:** a subfield of computer science that develops algorithms, which enable computers to learn from data to make improved decisions or actions.

10 **Sparsity:** the principle that most possible relationships between different variables will be of negligible size or importance.

11 **Multivariate data:** data for which more than one variable is being observed simultaneously.

12 **Occam's razor:** a principle devised by the 14th century friar William of Ockham, stating that of all possible solutions to a conundrum, the simplest should apply.

Sparse methods, and other advanced statistical techniques can be combined with machine learning models. Learning methods are normally assessed in terms of predictive performance (Breiman, 2001) (Tennenbaum et al., 2011), training algorithms on one set of the data, and testing their prognostic ability on another set. However, such methods may not capture the underlying factors influencing the values predicted, which are typically instructive for human understanding, interpretation and decision making. Flexible predictive models include regression trees[13], random forests[14], and deep learning[15], which are all techniques for predicting complex behaviour (Friedman et al., 2001). Deep learning was adopted recently by Deepmind to train computers to outclass humans at the board game *Go* (Silver et al., 2016).

Yet, most descriptive and predictive models, whilst highlighting trends or correlations, cannot define causal relationships, which are the direct dependence of particular variables upon others (Peters et al., 2015). The distinction between correlation and causation was highlighted by Lazer at al. (2014), in their discussion of Google Flu Trends, which aimed to predict outbreaks on the basis of related web searches, but significantly overestimated incidence. Attempts have been made recently to extend machine learning methods to recognise causality[16] (Shimizu et al., 2006) (Buhlmann & Hauser, 2012) (Janzing et al., 2014) (Lopez-Paz et al., 2015). One approach to studying causality is to manually force changes in one set of variables and monitor resulting changes in others. True causal relationships should withstand such intervention; otherwise the models would be pushed beyond their domain of applicability and may be expected to fail.

With in-built contributions from statistics, machine learning algorithms have great power for description, prediction and causal inference (Jordan & Mitchell, 2015) (LeCunn et al., 2015). As such, machine learning has been identified as a 'disruptive innovation', and is currently the focus of a Royal Society committee[17].

## Multiple testing

Even having identified correlations and causations, to usefully leverage data one must still consider the significance of any observed effects. Traditional significance tests[18] are a mainstay of scientific enquiry, but with large, complex datasets, one must apply such tests repeatedly, which increases the propensity that non-existent effects will be detected due to random chance. This is a recognised concern in domains such as neuroscience (Eklund et al., 2016), and it can mean that many studies are not reproducible (Baker, 2016).

To remedy this, innovative techniques are being adapted, such as the method of Benjamini & Hochberg (1996), to correct for multiple tests[19] and strike a balance between avoidance of false positives[20], and oversight of true effects. An emerging area, likely to grow in coming years, is selective inference (Taylor & Tibshirani, 2015). Selective inference recognises that often, multiple tests are run on the same data to investigate different effects, and corrects for this fact.

## New forms of analysis: advances in algorithms

Above, we described machine learning algorithms which use sparsity to identify the simplest explanations for phenomena. However, running such algorithms on large and complex data requires computationally-intensive storage and processing power. To overcome these bottlenecks, algorithms are increasingly optimised for efficient (Zou and Hastie, 2005) (Cevher et al., 2014), and stable (Meinshausen & Bühlmann, 2010) deployment, and increasingly run on distributed computing systems, which spread the computational burden across multiple machines. However, compartmentalising data and algorithms in this way can complicate analysis, so distributed algorithms require careful optimisation, a fascinating and growing research area (Boyd et al., 2011), (Arjevani & Shamir, 2015). The 'Bag of Little Bootstraps' paradigm, introduced by Kleiner et al. (2014), effectively balances

13 **Regression tree:** a method of prediction based on the probabilities of binary decisions.

14 **Random forest:** the use of many regression trees rather than a single tree for prediction.

15 **Deep learning:** an advanced method for representing and learning from complex data, which extensively models relationships between variables in a manner akin to neurones in the nervous system.

16 **Causality / causal models:** classical models indicate associations or correlations between variables. Causality moves beyond association to indicate interdependency (i.e. where alterations in one variable induce changes in another). Causal models typically embed a range of advanced statistical techniques (Pearl, 2010).

17 See **https://royalsociety.org/topics-policy/projects/machine-learning/**

18 **Significance test:** using statistical methods to determine whether an apparent difference between sets of observations can be considered genuine ('significant'), and thus whether it likely corresponds to a real effect, or is otherwise due to random chance or another factor. Classical significance tests posit a null hypothesis (that there is no difference), and an alternative hypothesis (that there is) and seek to determine whether the null hypothesis can confidently be rejected, thus pointing to the presence, or absence, of a true effect.

19 **Multiple test:** the implementation of more than one statistical significance test. Each individual test confers a small probability that the null hypothesis will be incorrectly rejected; performing multiple tests therefore compounds the likelihood that a non-existent effect will be detected. Multiple tests are needed to detect effects in large, complex datasets, but can result in erroneous deductions if not corrected for.

20 **False positive/ false negative:** where statistical tests incorrectly identify non-existent effects (false positive), or fail to detect true effects (false negative). Statistical tests are designed to fix the probability of a false positive detection as small and then minimize false negative detection.

**Figure 1: Magnetic resonance images of the brain**

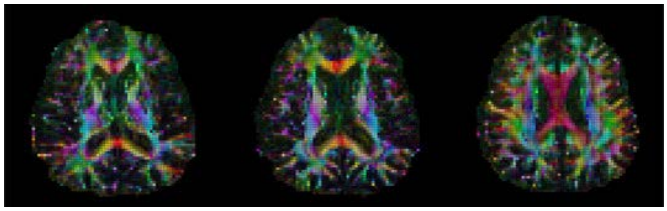


**Figure 2: Time series plots**

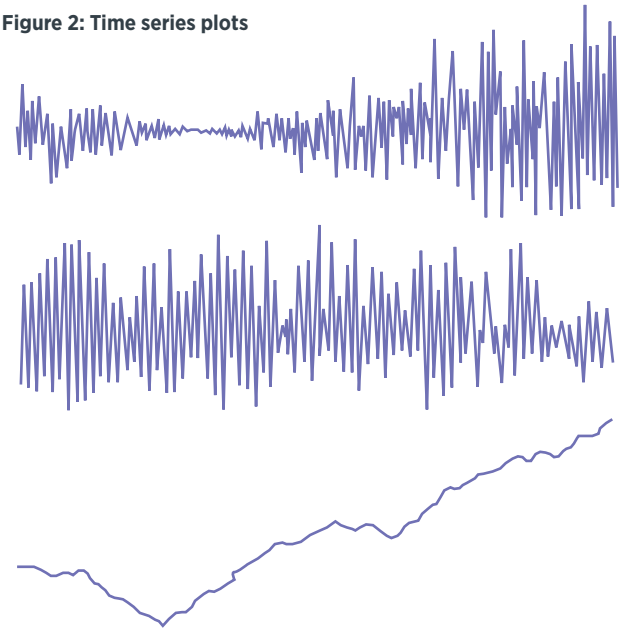


**Figure 3: Network data**

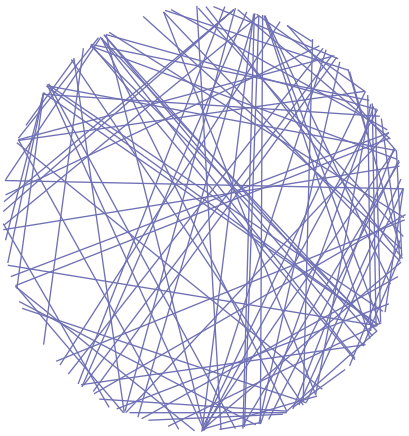
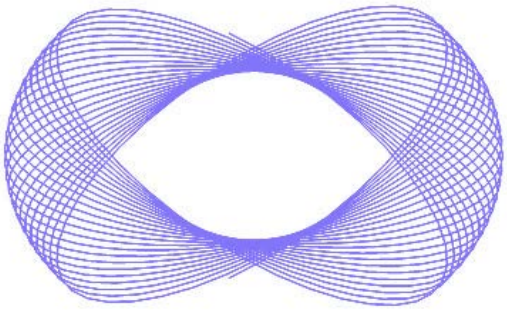


**Figure 4: Data mapped out as a shape in three-dimensional space**



computational demands with robustness and efficiency, and is an intriguing methodological approach.

The modern data analyst increasingly handles data of varying format and large dimensions. The emerging discipline of data science develops methods and techniques to extract useful information and insight from these data. This is presented in the following figures.

Figure 1 shows three diffusion-weighted magnetic resonance images of the brain. Each image is divided into smaller three-dimensional units called voxels, which are differentially coloured in proportion to the ability of water molecules to diffuse through each voxel. Irregularities in diffusion patterns can be indicative of disease, such as stroke and brain tumours. Medical images such as these therefore serve as important diagnostic and prognostic tools.

Figure 2 shows three time series plots, which chart the status of specific variables at successive points over time. This allows the behaviour of systems to be followed temporally, which enables the detection of variance and aberrant activity. Common examples of time series analysis are electroencephalography and electrocardiography, which describe electrical activity within the brain and heart respectively.

Figure 3 presents network data, depicting the structure of linkages, or relationships between entities within a system. This permits the identification of clusters of interconnection and activity, and can help the analyst determine more specifically the units or sub-groups within the system that are of greatest interest. Commercially, network analysis can be used for customer stratification and segmentation, to facilitate the targeted delivery of services.

Figure 4 depicts data mapped out as a shape in three-dimensional space. One example of such observations would be the tracking of an object's position over time.

The theoretical foundations of data science are evolving to encompass these and newer, emerging data forms.

New challenges are also arising for the ethical handling of data, in using algorithms that preserve privacy, and in determining causality between variables in observed data. This article outlines the current state-of-the-art in data science and some upcoming challenges for the field.

Further developments are also expected in algorithms that preserve privacy. As alluded to above, many public and private sector organisations hold personal data but cannot share it. Recent work in privacy-sustaining algorithms allows summaries to be compiled and inferences drawn from such data, at best without compromising personal information (Graepel et al., 2012), (Aslett et al., 2015), (Bos et al., 2014), (Kusner et al., 2015). As much of this data can yield important societal insight, such work will be of increasing relevance in the future.

## Discussion

When faced with a big data challenge, today's analyst can draw upon much theory developed over the past few decades, particularly in the areas of high-dimensional data analysis and machine learning (Donoho, 2015) (Fan et al., 2014). Yet, several challenges raised in this article require additional theory be developed for resolution.

At the outset, large raw datasets typically require pre-processing, cleaning and formatting, a suite of activities termed 'wrangling'. Though laborious, wrangling is essential to render analysis meaningful. However, the theoretical frameworks to maximise the consistency and efficiency of wrangling protocols remain poorly defined. New theory must also better correct for missing data and biased samples, which will enable 'found' data to be processed more consistently, and maximise rare event and subpopulation detection (Alyass et al., 2015) (Madigan et al., 2013) (Bühlmann & Meinshausen, 2016).

New theoretical developments are also required at the interface of predictive machine learning and statistical causal inference, to enable robust statistical principles to be scaled up via emerging computer platforms and algorithms, for application to large datasets. This is especially necessary for data in new formats and in streams.

Once data have been analysed, however, outputs must be presented in an accessible form for interpretation (Wolfe, 2013). New approaches for the visualisation of complex, high-dimensional data need to be developed, in order to clearly isolate groupings of interest, and facilitate decision-making.

Finally, the area of data ethics remains understudied. Transparent frameworks promoting public understanding of value and risks associated with sharing personal data need to be developed. Privacy, a related area, will see new theoretical approaches to enable analysis of confidential data with maintenance of anonymity. Further information on data ethics is given by Professor Luciano Floridi in section 9 of this Longevity Bulletin.

Data analysis has progressed markedly from the time of Tukey, but extracting useful and actionable insight still lies at its heart. New theoretical innovations will play a major role in tackling the challenges posed by big data, and realising the opportunities it offers.

## References


Alyass, A., Turcotte, M., Meyre, D. (2015). *From big data analysis to personalized medicine for all: challenges and opportunities.* BMC Medical Genomics 8: 33.

Arjevani, Y., and Shamir, O. (2015). *Communication complexity of distributed convex learning and optimization.* Advances in Neural Information Processing Systems, 1756-1764

Aslett, L.J.M., Esperança, P.M., Holmes, C.C. (2015). *A review of homomorphic encryption and software tools for encrypted statistical machine learning.* arXiv preprint arXiv:1508.06574.

Baker, M. (2016). *1,500 scientists lift the lid on reproducibility.* Nature 533(7604): 452-454.

Benjamini, Y., and Hochberg, Y. (1995). *Controlling the false discovery rate: a practical and powerful approach to multiple testing.* Journal of the Royal Statistical Society: Series B (Methodological), 289-300.

Bos, J.W., Lauter, K., Naehrig, M.(2014). *Private predictive analysis on encrypted medical data.* Journal of Biomedical Informatics 50: 234-243.

Boyd, S., Parikh, N., Chu, E., et al. (2011). *Distributed optimization and statistical learning via the alternating direction method of multipliers.* Foundations and Trends in Machine Learning 3(1): 1-122.

Brieman, L. (2001). *Statistical modelling: the two cultures.* (With comments and a rejoinder by the author). Statistical Science 16(3): 199-231.

Bühlmann, P., and Hauser, A. (2012). *Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs.* Journal of Machine Learning Research 13(August): 2409-2464.

Bühlmann, P, and Meinshausen, N. (2016). *Nagging: maximin aggregation for inhomogeneous large-scale data.* Proceedings of the IEEE 104: 126-135.

Cevher, V., Becker S., Schmidt M. (2014). *Convex optimization for big data: scalable, randomized, and parallel algorithms for big data analytics.* IEEE Signal Processing Magazine 31: 32-43.

Donoho, D. (2015). *50 years of data science.* Technical report, University of California Berkeley.

Efron B., Hastie T., Johnstone I., et al. (2004). *Least angle regression.* The Annals of statistics, 32(2): 407-499.

Eklund, A., Nichols, T.E., Knutsson, H. (2016). *Cluster failure: why fMRI inferences for spatial extent have inflated false-positive rates.* Proceedings of the National Academy of Sciences 113(28): 7900-5.

Fan, J., Han, F., Liu, H. (2014). *Challenges of big data analysis.* National Science Review 1(2): 293-314.

Friedman, J., Hastie, T., Tibshirani, R. (2001). *The elements of statistical learning.* (Springer series in Statistics): *volume 1.* Springer, Berlin.

Graepel, T., Lauter, K., Naehrig, M., (2012). *MIConfidential: machine learning on encrypted data.* In International Conference on Information Security and Cryptology, 1-21. Springer, Berlin.

Janzing, D., Chaves, R., Schoelkopf, B. (2015). *Algorithmic independence of initial condition and dynamical law in thermodynamics and causal inference.* arXiv preprint rXiv:1512.02057

Jordan, M.I., and Mitchell, T.M. (2015). *Machine learning: trends, perspectives, and prospects.* Science, 349(6245): 255-260.

Kleiner, A., Talwalkar, A., Sarkar, P., et al. (2014). *A scalable bootstrap for massive data.* Journal of the Royal Statistical Society: Series B 76(4): 795-816.

Kusner, M.J., Gardner, E.D.U.J.R., et al. (2015). *Differentially private Bayesian optimization.* arXiv preprint arXiv:1501.04080.

Lazer, D., Kennedy, R., King, G., et al. (2014). *The parable of Google flu: traps in Big Data analysis.* Science, 343: 1203-1205

LeCun, Y., Bengio, Y., Hinton, G. (2015). *Deep learning*. Nature 521(7553): 436-444

Lopez-Paz, D., Muandet, K., Scholkopf, et al. (2015). *Towards a learning theory of cause-effect inference.* Proceedings of the 32nd International Conference on Machine Learning, JMLR: W&CP, Lille, France, 2015.

Madigan, D., Ryan, P.B., Schuemie, M., et al. (2013). *Evaluating the impact of database heterogeneity on observational study results.* American Journal of Epidemiology, 178(4): 645-651

Manyika, J., Chui, M., Brown, B., et al. (2011). *Big data: The next frontier for innovation, competition, and productivity.* McKinsey Global Institute Report

Meinshausen, N., and Bühlmann, P. (2010). *Stability selection.* Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72(4): 417-473

Pearl, J. (2010). *An Introduction to Causal Inference.* The International Journal of Biostatistics 6 (2), 7

Pell, J.P., Valentine, J., Inskip, H. (2014). *One in 30 people in the UK take part in cohort studies.* The Lancet (London, England) 383(9922): 1015

Peters, J., Bühlmann, P., Meinshausen, N. (2015). *Causal inference using invariant prediction: identification and confidence intervals.* arXiv preprint arXiv:1501.01332

Royal Statistical Society (2016). *The Opportunities and Ethics of Big Data.* Available at **www.rss.org.uk/Images/PDF/influencing-change/2016/rss-report-opps-and-ethics-of-big-data-feb-2016.pdf** [Accessed 13/12/16]

Schimizu, S., Hoyer, P.O., Hyvärinen, A., et al. (2006). *A linear non-Gaussian acyclic model for causal discovery.* Journal of Machine Learning Research, 7(Oct): 2003-2030

Silver, D., Huang, A., Maddison, C.J., et al. (2016). *Mastering the game of Go with deep neural networks and tree search.* Nature, 529(7587): 484-489

Taylor, J., Tibshirani., R.J. (2015). *Statistical learning and selective inference.* Proceedings of the National Academy of Sciences 112(25): 7629-7634

Taylor, J.E., Worsley, K.J. (2008). *Random fields of multivariate test statistics with applications to shape analysis.* The Annals of Statistics 36: 1-27.

Tennenbaum, J.B., Kemp, C., Griffiths, T.L., et al. (2011). *How to grow a mind: statistics, structure, and abstraction.* Science 331(6022): 1279-1285.

Tukey, W. (1959). *The estimation of (power) spectra and related quantities.* In Langer, R. E. (1959) (ed.) On numerical approximation. University Wisconsin Press, Madison, WI., pp. 389-411.

Wolfe, P.J. (2013). *Making sense of big data.* Proceedings of the National Academy of Sciences, 110: 18031-18032.

Wu, X., Zhu, X., Wu, G.Q., et al. (2014). *Data mining with big data.* IEEE Transactions on Knowledge and Data Engineering 26: 97–107.

Zou, H., Hastie, T. (2005). *Regularization and variable selection via the elastic net.* Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67(2): 301-320

## Biographies

**Sofia Olhede** is a Professor of Statistics at University College London and holds an Honorary Chair in the UCL computer science department. Her research interests are big data, networks, non-stationary and non-linear time series and random fields, time-scale and time-frequency inference with applications in ecology, finance, and oceanography.

**Russell Rodrigues** is Operations Manager at the UCL Big Data Institute and Centre for Data Science. He manages the Institute's academic, corporate and governmental partnerships and coordinates its executive education programmes and conferences.

Follow UCL's data science activities on Twitter **@uclbdi**

# 4. Opinion: Is big data just big hype?

Professor Peter V. Coveney, University College London and Roger Highfield, Science Museum Group

With the rise of big data, there is a growing expectation that new insights into longevity will be deeper and richer than ever. Some have even concluded that we no longer have to understand biological processes at all, given the astonishing amount of genome data along with the rise of other "omes" – proteomes, microbiomes and transcriptomes. We can simply use machine learning to mine this trove of data for understanding lifespan. If only it were that simple.

Machine learning relies on a type of artificial intelligence, supposedly modelled on the brain, known as a neural network of processing units. These have been a subject of study since the 1940s but recently have become fashionable again with the rise of 'deep' networks containing large numbers of hidden layers of units.

But, as we have argued recently in a paper (Coveney et al., 2016) with a colleague, the more extravagant claims made for the power of big data alone rest on shaky foundations: big data demand big theory too.

No matter how deep or sophisticated they are, neural nets merely fit curves to existing data. In effect, they say "based on all the people we have examined before, we can predict your risk of disease and life expectancy." Although we know they have grown in power in the past thirty years, there is no rigorous theory to explain how well they work.

They typically fail in circumstances beyond the range of the data used to train them. As the overestimate of peak influenza levels by Google Flu Trends showed, past success in describing epidemics is no guarantee of future performance.

Paradoxically, the bigger the data, the more likely we are to swamp the details of how an individual person will respond. Because we are all different, the only way to use genetic information to predict how long a person may live is if we have a profound understanding of how the body works, so we can model the way that a particular person will age.

Because we lack the understanding to do this, the next best thing is to look at how genetically similar people react and then assume that a given person will respond in a similar way – that is why people talk about 'precision medicine' (even though it is less precise than personalised medicine!)

But there are many other issues with blindly applying big data: without understanding to guide the collection and curation of data, there are many potential pitfalls because, in biology, big data is actually tiny relative to the complexity of a cell, organ or body. Few give much thought to how the body is a dynamic system, subject to cycles, circadian rhythms and constant renewal.

One needs theory to help understand which data are important for a particular objective. When it came to the discovery of the Higgs particle at the Large Hadron Collider in Geneva, for example, the gathering of petabytes of data was guided by theory developed decades ago. Nor do we predict tomorrow's weather by averaging historic records of that particular day's weather – mathematical models do a much better job with the help of daily data from satellites. Similarly, trying to forecast a patient's lifespan based on thousands of others is like trying to forecast the weather on a given date by averaging historic records of that day's weather.

We have to be sceptical about the data that we gather. The fact that "most published research findings are false", as famously reported by John Ioannidis from Stanford University, underlines that one critical dataset – the conclusions of peer reviewed studies – is unreliable without good experimental design and rigorous statistical analysis. Quality is more important than quantity.

We have to be sceptical about the range of data we gather. Most assume that "Big" means "lots and lots of data points"; Xiao-Li Meng at Harvard University has demonstrated that one can only make reliable predictions from big data if they truly represent a big fraction of the actual population of interest.

We have to be sceptical about correlations. As the old saying goes, correlation is not causation and spurious correlations (the 'clever Hans effect', for example) are a familiar headache for anyone who has tried to use machine learning to predict the biological activity of molecules.

The bottom line is that even bigger data are not enough. To effectively use the explosion in big data, we need to improve the modelling of biological processes. We need models to understand the sensitivity of complex biological systems to tiny errors in data. In high dimensional spaces, where we are very unlikely to be able to harvest enough data to make rigorous inferences using machine learning, we need underlying theory and understanding in order to reduce of amount of data required in the first place.

We can only move from simplistic look-up tables of cause and effect to true science when we have understanding that can provide reliable insights in novel circumstances. We need models that are truly predictive. They have to be actionable, not only in the weak sense that they can be used post hoc, for instance to help hone a drug action or process, but in a strong sense that they can be used to predict the future so that action can be taken before it becomes a reality, as is already the case when forecasting severe weather.

In medicine, the most vivid example of an actionable prediction is one that can extend the life of a patient. That could, for example, mean a prediction that enables a doctor to pick one antimicrobial drug in preference to another when confronted with a severe infection. One author of this article (Peter V. Coveney) is already reporting results that show how it will soon be possible to take a person's genetic makeup and – with the help of sophisticated modelling, heavyweight computing and clever statistics – select the right customised drug in a matter of hours.

This is why the European Commission is investing five million Euros in the CompBioMed (**www.compbiomed.eu**) initiative led by Peter Coveney's team at University College London (UCL). True understanding of the factors affecting human longevity will arise from this kind of approach, not from blindly groping around for correlations in vast datasets harvested from complex biological systems. One day it should be possible use big data with modelling to create virtual humans, so a person's digital Doppelganger can provide a glimpse of what is in store, from the effect of treatments to the impact of diet on lifespan.

## References


Butler D. (2013). *When Google got flu wrong.* Nature 494: 155-156. Available at: **https://www.youtube.com/watch?v=8YLdIDOMEZs**

Coveney P.V., Dougherty E.R., Highfield R.R. (2016). *Big data need big theory too.* Philosophical Transactions of the Royal Society A 374, 20160153. Available at: **http://rsta.royalsocietypublishing.org/content/374/2080/20160153**

Coveney, P. V. and Wan, S. (2016). *On the calculation of equilibrium thermodynamic properties from molecular dynamics,* Physical Chemistry Chemical Physics, 2016, 18, 30236-30240, DOI:10.1039/C6CP02349E

Meng, X.L. (2014). *A Trio of Inference Problems that Could Win You a Nobel Prize in Statistics (If You Help Fund It).* In Past, Present, and Future of Statistical Science (Eds: X. Lin, et. al), CRC Press, pp. 537-562.


## Biographies

**Peter Coveney** is Director of the Centre for Computational Science, Professor of Physical Chemistry, and an Honorary Professor in Computer Science at University College London. He is Professor Adjunct at Yale University School of Medicine (USA). He is active in a broad area of interdisciplinary research including condensed matter physics and chemistry, materials science, as well as life and medical sciences in all of which data-intensive high performance computing plays a major role.

**Roger Highfield** is the Director of External Affairs at the Science Museum Group. He was the Science Editor of The Daily Telegraph for two decades and the Editor of New Scientist between 2008 and 2011. With Peter Coveney, he wrote the bestseller, The Arrow of Time: The Quest to Solve Science's Greatest Mystery, and Frontiers of Complexity: The Search for Order in a Chaotic World.

# 5. Routes to diagnosis for cancer

Lucy Elliss-Brookes, Head of Cancer Analysis, National Cancer Registration and Analysis Service, Public Health England

First published in 2010, Routes to Diagnosis used a novel methodology to harness and exploit large cancer datasets, identifying the route a patient took through the healthcare system before receiving a cancer diagnosis. Unexpected differences in how patients were diagnosed were uncovered, including large variation in short-term survival and many inequalities across different patient groups and cancers. It has been a major driver behind a national and international focus on early diagnosis for people with cancer. Understanding diagnostic pathways and our ability to influence them are now an important part of tackling cancer, with a reduction in emergency presentations being an important aspect of this. Updates have been used to chart the impact of early diagnosis campaigns, improved treatments and the evolution of national screening programmes. This ground-breaking project filled a large knowledge gap to the benefit of cancer patients. Results are used to monitor the changes in the distribution of cancers, and to understand better where we can best focus our efforts to improve outcomes.

## Introduction

An important pillar of recent national cancer strategies is the promotion of early diagnosis of cancer, thereby improving survival rates and reducing cancer mortality. In order to do this we needed to understand how patients are diagnosed with cancer. We knew that survival rates for cancer in England and the UK were poor compared to our European counterparts and suspected it might be due to later diagnosis.

In 2009 we knew that less than 10% of cancers were diagnosed through screening, and understood something about the percentage of people being referred as a Two-Week Wait[22] (TWW), but we had no idea how many came via other GP referrals, or through Accident & Emergency (A&E), or were picked up in secondary care, say when a patient is being treated for an unrelated condition. The suspicion was that a significant amount of late cancer diagnoses arise in these cases where patients have not gone through a 'managed' route. There was speculation that we may find out that as many people are diagnosed through going to A&E as are diagnosed through all our current screening programmes put together - quite a sobering thought.

The challenge was to use routine datasets and consider how we could mesh together a variety of data sources to understand patients' routes to diagnosis. The intention was to identify a route for all cancer patients, not just those of 'the big four'. Then the results could be scrutinised by route, age, sex, ethnicity, deprivation and geographical area. Crucially, the patient outcomes, namely survival time after diagnosis, could be examined and compared.

## Approach

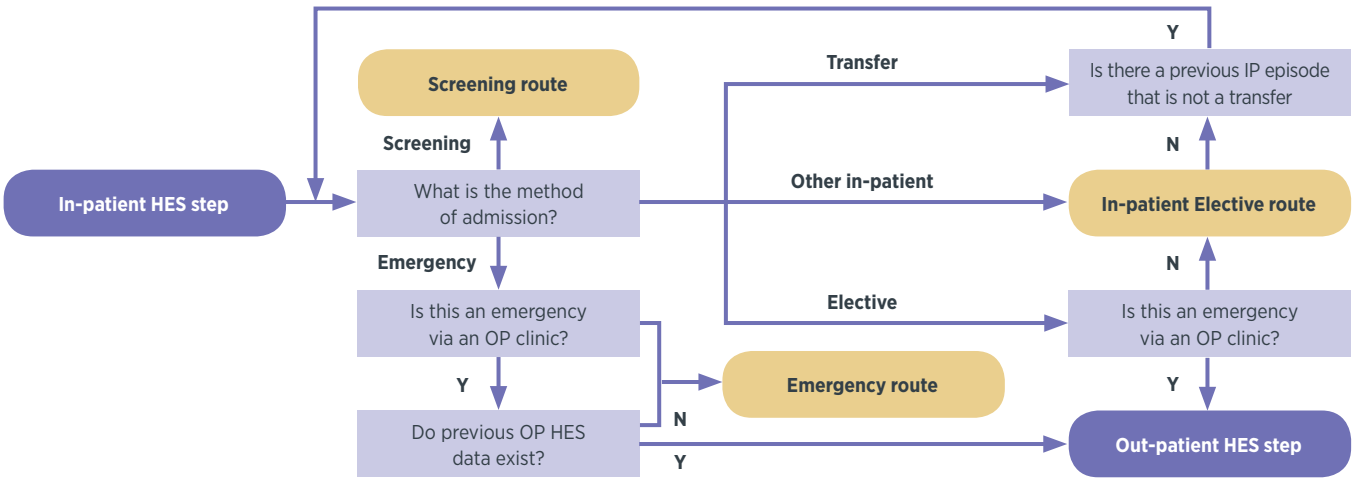
The approach taken was to undertake large-scale linking of national datasets with cancer registration data. This enabled comprehensive coverage though was subject to information governance issues and dataset availability.

Datasets were linked at tumour level using NHS number. The algorithm first used Hospital Episodes Statistics (HES) data to categorise the route for each tumour individually. The project team identified 135 different pathways to diagnosis; these were grouped into eight Route categories. National Screening Programme and Cancer Waiting Times (CWT) data linked by NHS number to the cancer registration record were examined with the assignment of route potentially changing to either a 'Screening' or TWW route. For cases with no HES activity the route was classified as Unknown or Death Certificate Only (DCO).

Detailed flow diagrams were drawn up illustrating the categorisation of patients into routes. Figure 5 shows the steps taken to seek a start point to the route when the end point was an inpatient admission.

[22] Two-Week Wait is a rapid referral route whereby patients being urgently referred for suspected cancer by their GP can expect to be seen by a specialist within two weeks.

**Figure 5: Flow diagram for finding the start point or prior step for an in-patient step in a route**




In-patient (IP): a patient who is admitted to hospital for a procedure or diagnosis
Out-patient (OP): a patient who attends a healthcare appointment without staying overnight

## Results

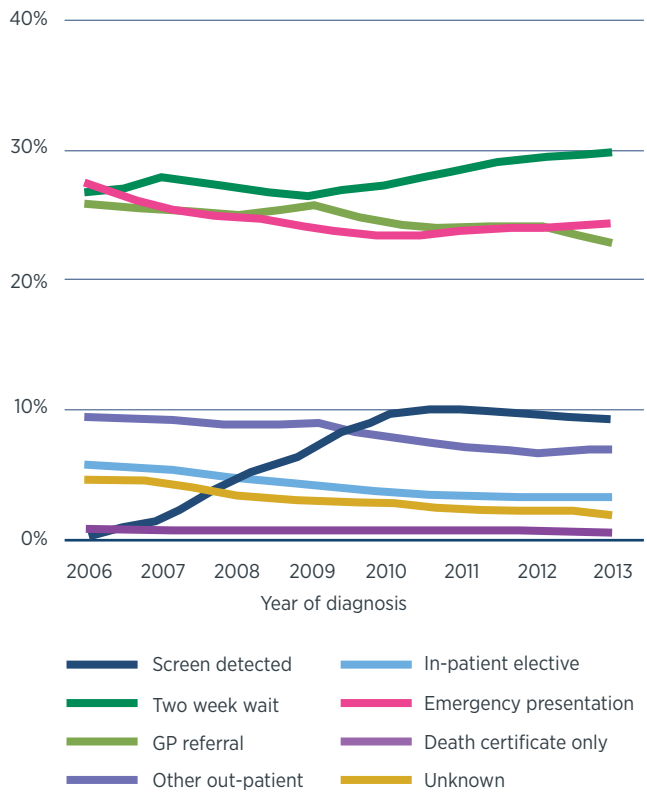
The initial publication revealed for the first time the proportion of cancers diagnosed as an emergency presentation – one in four cancers, and that the survival for this cohort was the lowest out of those analysed. The most recent data covering 2006-2013 show a reduction in these emergency presentations, down to 20% for all cancers and an increase in the TWW referral route. The results also cover the introduction and roll-out of the bowel cancer screening programme, with an initial rise in the proportion of screen-detected cancers seen for the relevant age ranges, plateauing at around 27% for 60-69 year olds. Survival for this route is high, a trend seen for the other national screening programmes (breast and cervical).

Results showed a large variation by cancer site, with 56 cancers included in the latest publications. This focus on the less common cancers can be used to inform site specific work and awareness campaigns, as well as to support the vital work of the smaller cancer charities and patient groups. Variation was also seen between the sexes for some cancers, but more striking was the variation by age – with older age groups having a high proportion of emergency presentations. Inequalities by deprivation were also seen for some sites.

Our motivation lies within the rich detail that the results are able to give us, and in the variation and inequalities that our data continue to reveal. When we look behind the averages we are confronted with stark realities, for example:

- 50% of pancreatic cancer patients present through an emergency route, with only 10% of those people surviving a year after diagnosis.
- For women diagnosed with ovarian cancer, 20% of those under 50 are diagnosed as an emergency, compared to 51% of those aged over 85.
- For people with colorectal cancers, only 7% in the most deprived areas are diagnosed through screening compared to 11% in the least deprived areas. The proportion of emergency presentations has decreased between 2006 and 2013 and the proportion via TWW has increased as shown in Graph 1.

**Graph 1: Percentage of diagnoses by presentation route, Colorectal, by year**



In big data terms, over 118 million records were used to generate the latest set of results; more records are being used to produce the next update.

Many high profile publications that have used the Routes to Diagnosis results, including reports and articles from cancer charities, parliamentary groups, ministers, commissioners and clinicians: "One of the most shocking statistics is that a quarter of UK cancers are first detected at hospital accident and emergency departments" - John Baron MP, House of Commons, London, UK (published in *The Lancet: Oncology,* January 2012).

The results have also received media coverage, as well as being referenced and cited extensively in academic publications. A number of additional outputs have been produced by the team including posters and data briefings. Research currently being prepared for publication includes examination of prior GP consultations among cancer patients diagnosed as emergency presentations, and examination of the change in survival for different routes over time.

## Conclusion

The novel approach was subject to in-depth scrutiny by the cancer intelligence community following the publication of the initial results in 2010. Subsequently the methodology was published in *British Journal of Cancer* in 2012. Each updated publication of Routes to Diagnosis results has been the result of a review of the approach and the methodology, and improvements have been made wherever possible. The data underlying the work has also improved, as evidenced by the reduction over time of those assigned to the unknown route.

The Routes to Diagnosis study has formed the basis of a large number of other academic studies, some of which have examined and inspected the methodology in detail. The 2012 paper has been cited 120 times in the scholarly literature, (Google Scholar as at 06 December 2016), along with numerous other reports and publications. The data briefings and workbooks have also been frequently referenced.

Thanks to the integration of the work into Public Health England's national Cancer Analysis System (CAS) it has been possible to utilise the Routes to Diagnosis results easily in other work. Results are stored at patient and tumour level in a secure environment, with access to the data regulated by Section 251 of the NHS Act [2006]. Routes have been linked to data on surgical resections, further linkages with treatment data are now possible, including with radiotherapy and chemotherapy data. This would enrich our understanding of the issues associated with access to potentially curative treatment.

The improved staging information in the CAS allowed the possibility of extending the analysis on Routes to Diagnosis to include cancer stage; this was then expanded to look at variation in Routes by ethnicity. Routes to Diagnosis results also provide a key metric being used in the evaluation of past and current Be Clear on Cancer campaigns, which aim to help patient spot symptoms of cancer earlier.

Too many people with cancer are still being diagnosed too late. We have a responsibility to utilise the cancer data that we are entrusted with to ensure that we are doing all that we can to help understand why this is and what can be done to make a difference to patients. Routes to Diagnosis is one small but significant part of this endeavour, and one that I am immensely proud to have been involved in for the last seven years.

## References


Elliss-Brookes, L., McPhail, S., Ives, A., Greenslade, M., Shelton, J., Hiom S., Richards, M. (2012) *Routes to diagnosis for cancer – determining the patient journey using multiple routine data sets.* British Journal of Cancer 107(8): 1220-1226.

McPhail S., Elliss-Brookes L., Shelton J., Ives A., Greenslade M., Vernon S., Morris E.J., Richards M. (2013). *Emergency presentation of cancer and short-term mortality.* British Journal of Cancer 109(8): 2027-2034.

Abel, G., Shelton, J., Johnson, S., Elliss-Brookes, L., Lyratzopoulos, G. (2015). *Cancer-specific variation in emergency presentation by sex, age and deprivation across 27 common and rarer cancers.* British Journal of Cancer 112: S129-S136.

Zhou, Y., Abel, G., Hamilton, W., et al. (2016) *Diagnosis of cancer as an emergency: a critical review of current evidence.* Nature Reviews Clinical Oncology


## Biography

**Lucy Elliss-Brookes** is the Head of Cancer Analysis at the National Cancer Registration and Analysis Service in Public Health England. She currently leads a team of highly skilled cancer analysts and managers responsible for measuring cancer outcomes across England and the UK and supporting national public health and healthcare policies aimed at reducing mortality from cancer.

# 6. Use of big health and actuarial data for understanding longevity and morbidity risk

Professor Elena Kulinskaya and Lisanne Gitsels, PhD candidate, University of East Anglia

## Introduction

Estimating longevity risk and evaluating associated uncertainty is one of the main topics of concern to actuarial community. It is well known that longevity is increasing considerably both in developed and developing countries, including the United Kingdom. We believe that to be able to establish the drivers of this change, and to predict how they may change over time and how this would affect life expectancy, researchers need to harvest Big Health Data (Hemmingway, 2014), i.e. to access large health databases, and to use sophisticated tools for modelling the mortality experience of participating populations using individual level health data. Big Actuarial Data such as the Continuous Mortality Investigation (CMI) data are of the utmost importance in translating the results to the reference population of relevance to the actuarial community.

Contemporary evidence-based underwriting needs to account for a large number of important and time-varying determinants of health and longevity, such as demographic factors (gender, social class), lifestyle factors (smoking, obesity, alcohol usage) and medical advances, and their interactions. Many public health interventions are aimed at increasing the health of populations. These vary from offering flu vaccination to encouraging lifestyle changes to management of chronic medical conditions. However, actuarial and medical research often aim at somewhat differing objectives. While mortality is of primary interest to an actuary, exacerbation of medical conditions is often the interest of a medical researcher. For instance, not death but a cardiac event may be the primary endpoint in many medical studies of heart disease or smoking. Additionally, clinical trials while of the gold standard when studying medical interventions, deal with a selective population of patients, and usually are of short duration.

This explains why the existing medical publications and their syntheses published in numerous systematic reviews, though certainly important, are not sufficient for actuarial purposes, and the direct involvement of actuarial researchers in the modelling of health-related data is of utmost importance. In-depth actuarial longevity research should concentrate on statistical modelling of population-based individual level data collected over the long term. Some advances in this direction are already being made (Ryan et al., 2013; Lu et al., 2014).

The title of this article (Use of big health and actuarial data for understanding longevity and morbidity risks) is, in fact, the name of a research programme recently funded by Institute and Faculty of Actuaries (IFoA). This is a joint project between the School of Computing Sciences and Norwich Medical School within the UEA, and Aviva Life. This research will use the data on 3.4 million patients born before 1960 from The Health Improvement Network (THIN) primary care database, and also the CMI data. The main objectives are the development of novel statistical and actuarial methods for modelling mortality, modelling trends in morbidity, assessing basis risk and evaluating longevity improvements based on individual level big health and actuarial data.

## Programme description

The first aim of the programme is the mortality modelling. This includes identification and quantification of the key factors affecting mortality/longevity such as lifestyle choices, medical conditions and/or interventions. A target list will include between 3-5 conditions or interventions. Statin prescription, an established longevity-improving intervention (Longevity Science Panel, 2014) is one of the target scenarios. The choice of the medical and social developments to be included in

research will be based on current models of disease burden in England (Newton et al., 2013), and combined with the availability of relevant information in the general practice data such as THIN.

The top causes of premature death in England are: heart disease, stroke, respiratory disease, cancer and Alzheimer's disease. Important health interventions and social developments include widening of statins prescription, possible changes in blood pressure targets, rise in obesity and type 2 diabetes, reduction in smoking, trends in diet and physical activity according to socioeconomic status. Some of the information required for tackling these diseases and interventions is not available in THIN. This includes the details on cancer severity, and this, unfortunately, takes cancer off the list.

After careful consideration of the importance of various conditions, interventions and lifestyle factors, and availability of the required information, the research team agreed on the following list which includes the main cardiovascular conditions: myocardial infarction, heart failure, atrial fibrillation, and stroke. The lifestyle factors of interest in respect to cardiovascular disease are smoking and obesity. Additionally, type 2 diabetes would contribute to all of the above conditions. Health interventions to study include statin prescriptions and a possible change in systolic blood pressure (BP) targets to 120 mm Hg. This is a very novel possible development, following the results from the just published SPRINT trial (2015). In this large trial, the lower BP target resulted in considerably lower all-cause mortality (hazard ratio, 0.73; 95% CI, 0.60 to 0.90). However this may also bring side effects, such as rise in acute kidney injury.

The second aim is the modelling of trends in morbidity and the uptake of health interventions. Trends in the incidence and/or prevalence of particular medical conditions and/or lifestyle factors will also be obtained from the primary care data, establishing patterns due to social or geographic inequalities, such as socio-economic status (SES), age or postcode lottery. For instance, the patients in the more deprived areas may be disadvantaged in regards to the latest interventions. A new intervention may be of benefit to only the most privileged individuals, at least initially. Similarly, outcomes of a public-health campaign aimed at healthier lifestyle choices are often associated with SES and will, therefore, result in SES dependent changes in the incidence of a disease. This will l ead to widening the gap in longevity between individuals from different backgrounds. Thus to be able to ascertain an effect on longevity of a population, the incidence of a condition or an uptake of an intervention needs to be modelled over time in parallel to modelling mortality.

As often happens with an existing portfolio of insured lives, the precise health details of a life are not available. Instead, the interest lies in the mortality trends of the whole book. To be able to provide this information, three components are required:

- established in Aim 1 model for survival differentials associated with a particular disease or intervention;
- developed in Aim 2 model for the incidence/prevalence of this condition or uptake of this intervention over time,
- the sufficient knowledge of the population to which it is desired to translate trends in longevity established in general population to be able to assess the basis risk (Haberman et al., 2014). The data submitted to the Continuing Mortality Investigation will be used for this purpose.

Finally, an open source R package will be developed. It will incorporate the models derived from the analyses of THIN and CMI data and provide analytical and graphical means to forecast longevity of a general UK population, and also of a population of a user defined composition under a number of scenarios for changes in disease incidence, health behaviors and treatments. This will be an open source software available from the project website along with an accompanying manual for its use. Teaching materials for the actuarial community on the modelling techniques used in the project, and the use of the developed R package will be available from the project website.

Our programme is funded by IFoA for four years from October 2016. However, we expect to obtain the first results and to present them to actuarial community within the first year.

## A case study: statins and longevity

This case study focuses on longevity improvement due to the widening guidelines on the prescription of statins to healthy patients. The results below are based on the preliminary research within Aim 1 by the second author, and are published in Gitsels et al. (2016).

Cardiovascular disease (CVD) is one of the main causes of death, accounting for 28% of all deaths in the United Kingdom. Statins are prescribed for primary and secondary prevention of CVD. For primary prevention, the risk of CVD is quantified by the so called QRISK2 score as the 10-year risk of a first cardiac event. In July 2014, the National Institute for Health and Clinical Excellence (NICE) lowered the risk threshold for which statins are prescribed at from 20% (2006 recommendation) to 10% (NICE, 2014). This translates to an increasing number of people being eligible for the drugs; that is an additional 4.5 million UK residents. From an actuarial perspective the question becomes whether the new NICE policy would materially affect mortality in the UK, and if yes then how.

The objective of our study was to estimate the survival benefits of statins for different risk groups at various ages in the general population. Data from THIN database were used, comprising medical records from 1987 to 2011 of people born between 1920 and 1940. Four cohorts aged 60, 65, 70, or 75 years with no previous history of CVD were studied, with sample sizes 118,700, 199,574, 247,149, and 194,085, respectively.

The hazard of mortality associated with statin prescription in patients at <10%, 10-19%, or ≥20% CVD risk was calculated by a multilevel Cox proportional hazard regression, adjusted for covariates including sex, year of birth, Mosaic (lifestyle groups defined by postcode), diabetes, blood-pressure regulating drugs, high cholesterol, Body Mass Index (BMI), and smoking status.

There was low uptake of statin therapy in the eligible population as seen in Graph 2. People at <10% CVD risk did not have a mortality benefit from statin prescription at any age, whereas people at 10-19% CVD risk had a mortality benefit of 11-21% by the age of 70. Furthermore people at ≥20% CVD risk had a mortality benefit of 14-18% by the age of 65 as shown in Figure 6.

**Graph 2: Statins prescriptions rates in the UK based on the THIN data**

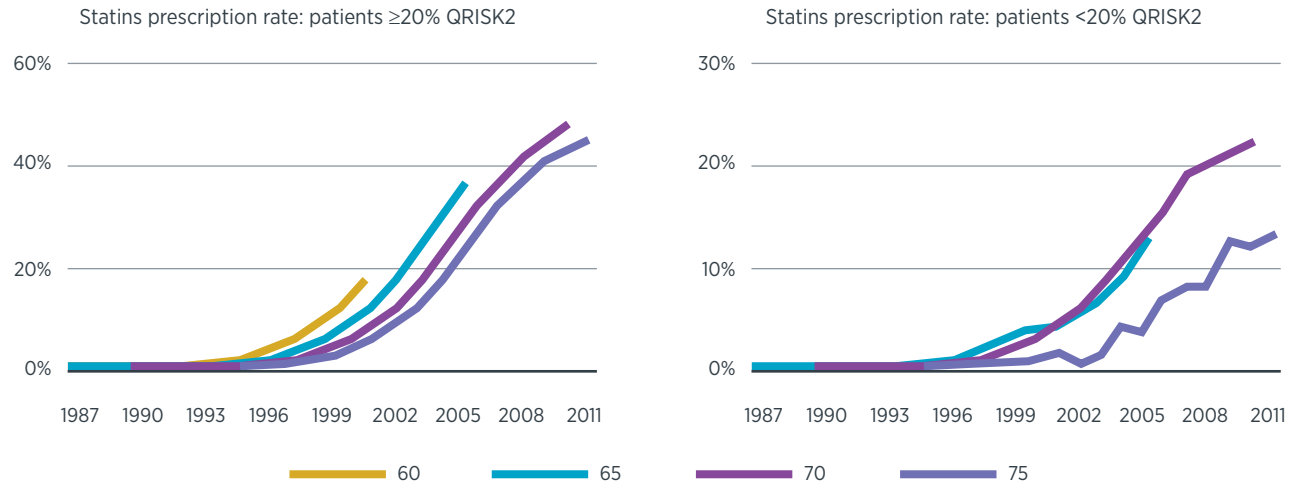
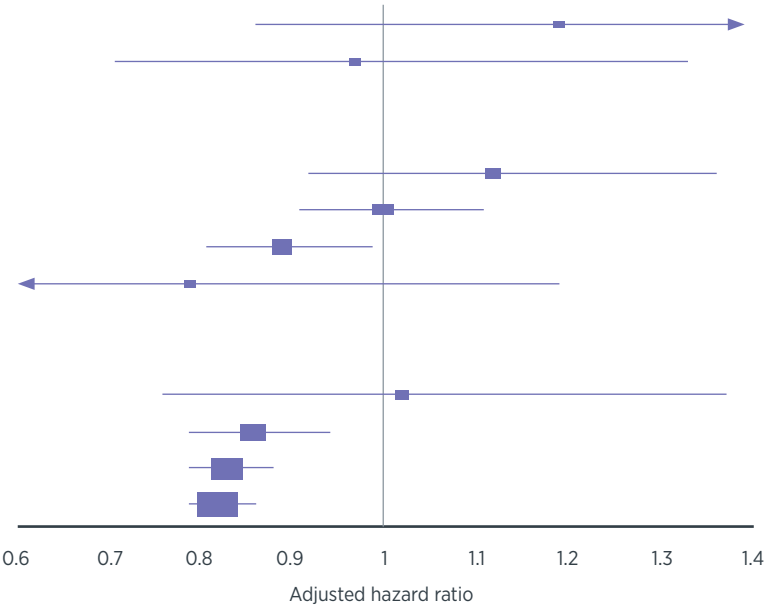



**Figure 6: Hazard ratio of death given statins prescription for patients stratified by QRISK2.**

**Hazard ratios adjusted for sex, year of birth, socioeconomic status, diabetes, hypercholesterolaemia, blood pressure regulating drugs, body mass index, smoking status, and general practice.**

| Cardiac risk at baseline | Adjusted HR (95%CI) |
|---|---|
| **<10%** | |
| Age 60 | 1.19 (0.86 - 1.65) |
| Age 65 | 0.97 (0.71 - 1.33) |
| **10 - 19%** | |
| Age 60 | 1.12 (0.92 - 1.36) |
| Age 65 | 1.00 (0.91 - 1.11) |
| Age 70 | 0.89 (0.81 - 0.99) |
| Age 75 | 0.79 (0.52 - 1.19) |
| **>=20%** | |
| Age 60 | 1.02 (0.76 - 1.37) |
| Age 65 | 0.86 (0.79 - 0.94) |
| Age 70 | 0.83 (0.79 - 0.88) |
| Age 75 | 0.82 (0.79 - 0.86) |



The mortality benefits translate to an increase in life expectancy of 1.2 to 2 years, respectively. In the course of the project we shall extend these results and combine them with a novel model for the uptake of statins over time (Aim 2) and we will develop an adjustment for the basis risk based on the CMI data (Aim 3) to provide a plausible scenario of temporal changes in longevity due to statins. Our model can also be used for the future cost-benefit analysis of the new NICE guidelines which would account for the additional drug costs and additional healthcare resource use.

## References


Haberman S. et al. (2014). *Longevity basis risk: a methodology for assessing basis risk.* Research report by Cass Business School and Hymans Robertson LLP

Hemmingway, H. (2014). *Big Health Data.* International Mortality and Longevity Symposium, Birmingham 2014.

Lu, J. L. C., Wong, W., Bajekal, M. (2014). *Mortality improvement by socio-economic circumstances in England* (1982 to 2006). British Actuarial Journal 19: 1-35.

National Institute for Health and Care Excellence (2014). *Wider use of statins could cut deaths from heart disease.*

Newton J.N. et al. (2013). *Changes in health in England, with analysis by English regions and areas of deprivation, 1990–2013: a systematic analysis for the Global Burden of Disease Study,* The Lancet 386(10010): 2257-2274.

Ryan, D., Rion, S., Rechfeld, F. (2013) *Disease-based models of longevity: the future of human longevity: cardiovascular health, longer lives.* 10-12 November 2013. Swiss Re.

Longevity Science Panel (2014). *What is aging? Can we delay it?*

The SPRINT Research Group (2015). *A randomized trial of intensive versus standard blood-pressure control.* New England Journal of Medicine , 373:2103-2116

Gitsels L.A., Kulinskaya E., Steel N. (2016). *Survival benefits of statins for primary prevention: a cohort study.* PLoS ONE 11(11): e0166847. doi:10.1371/journal.pone.0166847


## Biographies

**Elena Kulinskaya** is Professor in Statistics (Aviva Chair in Statistics) at UEA. Her research interests are meta-analysis and evidence synthesis, statistical methods of big data analysis, actuarial and medical statistics and design of statistical software.

**Lisanne Gitsels** is a PhD candidate and Senior Research Associate at UEA. Her research interests are in analysing longitudinal data to contribute to solving current social, health problems from an epidemiological and economic point of view.

Actuarial Research Centre®
Institute and Faculty of Actuaries

The Actuarial Research Centre (ARC) is an international network of actuarial researchers around the world delivering cutting-edge research programmes that aim to address some of the significant, global challenges in actuarial science. To find out more about this programme and other world-class research programmes, please visit: **www.actuaries.org.uk/ARC**

# 7. Personalised risk prediction: genomics and beyond

Dr Philippa Brice, PHG Foundation

The ability to predict more accurately the risk of diseases in individuals and within populations should form an important element in the practice of truly personalised medicine, providing opportunities to prevent or limit disease. Genomic and other forms of biological data can feed into risk prediction models to refine and improve individual risk assessments. Whilst the enormous and as yet poorly understood complexity of human biology and genomics significantly limits this capacity at present, there are already some opportunities for better prediction and prevention of ill-health. Combining real-time data from both biological and environmental sources, facilitated by emerging health technologies and capacity for big data analysis, may ultimately offer the most personalised solutions.

The future of personalised medicine will certainly include more precise risk prediction, informed by genomics, environmental and behavioural factors, physiological and other biological measurements. Eventually, we may progress to a state where risk prediction forms just one part of a continuous lifelong process of monitoring and maintaining health – but there is still a long way to go before we get there.

## Introduction

No two individuals are exactly alike (not even identical twins), and there is increasing recognition that a move away from 'one size fits all' healthcare to a more tailored approach offers benefits. An important component of personalised healthcare is the capacity to identify those in whom the risk of disease is substantially increased, whether at the individual level or as a population sub-group. This enhances the potential for preventive interventions, as well as for increased screening, earlier diagnosis and prompt treatment that can cure or at least reduce the severity of disease – though it may also increase the risks of 'over-diagnosis' and unnecessary treatment.

Broadly speaking, health and disease arise from highly complex interactions of innate (genetic) and external (environmental and behavioural) factors. In rare inherited diseases, genetic influences dominate, but in most cases the picture is more complicated. Genetics can influence our behaviour and modulate environmental interactions; conversely, external agents can influence the regulation of our genes. Whilst the effects of 'lifestyle' factors such as diet and exercise on the risk of conditions like heart disease or cancer are well known, making sense of the interactions between genetic and environmental factors in disease risk and causation is much harder.

Nevertheless, genomics (underpinned by rapid developments in DNA sequencing and analysis) offers new opportunities for risk prediction. At the same time, parallel developments in a host of other scientific fields and technologies are enabling improvements in health monitoring, including via the use of 'wearables' and biological sensors. So as we approach what is being widely hailed as the 'era of personalised medicine', how far can our capacity to assess risk and predict disease improve?

## Identifying high-risk groups from within populations using genomics

It is increasingly feasible to identify rare disease-associated genetic mutations in individuals and their families. Thousands of rare inherited conditions (such as muscular dystrophy, cystic fibrosis or Marfan Syndrome) are now known, and added to these are 'new' (de novo) genetic changes such as the chromosomal abnormalities that cause Down Syndrome or Fragile X Syndrome. Collectively, one in 17 people in the UK have a rare disease; genetic testing can pinpoint the precise cause of such conditions and inform clinical management. It can also help prevent disease and even death in family members who share the same genetic mutation.

For example, identifying familial hypercholesterolaemia by DNA testing in someone who has had a heart attack in early adulthood means that their children can also be tested. Affected children may show no signs of disease, but should nevertheless receive special care to keep their own cholesterol levels low and prevent potentially fatal premature cardiovascular disease. Similarly, families affected by hereditary breast-ovarian cancer syndrome caused by BRCA gene mutations can be tested; women without a mutation have only population-level cancer risk. Those with it have greatly increased risk – but this knowledge at least allows them to opt for earlier and more frequent screening, or even for full surgical removal of the breasts and ovaries.

Adults with siblings or children affected by genetic diseases can learn the risks of future children having the same disease, and the options available to avoid this, such as pre-implantation genetic diagnosis (PGD), whereby IVF embryos are screened to select those without the disease mutation.

## Common diseases and risk stratification

Genetic diseases are unusual in being primarily (if not exclusively) caused by genetic changes – though even then, the probability of clinical disease and its severity when it occurs may vary. Conversely, common complex diseases arise from a combination of multiple contributory environmental and genetic factors. Genetic variants that can contribute to overall risk of a given disease in an individual are typically relatively common in the population, and individually most confer very modest changes in risk. However, their cumulative effect may be greater, creating a spectrum of genetic risk across different individuals within a population.

There is evidence that incorporating genetic data on the presence or absence of multiple mutations into risk prediction algorithms (along with usual data on age and sex) could refine and improve risk assessment. This might, for example, identify some women at relatively high risk of breast cancer who should be offered screening, but would be excluded from current programmes because they are too young. Improved, stratified risk prediction is feasible in the not-too-distant future for at least some common conditions, and could be relevant for insurance as well as public health.

## Wider opportunities to identify risk and prevent or limit ill-health

Specific gene-environment interactions can play a significant role in disease. For example, tobacco smoking is strongly associated with an increased risk of lung cancer, but not all smokers are affected to the same degree. Studies in different populations have identified a host of genetic variants that appear to influence physiological interactions with tobacco smoke to affect lung cancer susceptibility. The field of 'toxico-genomics' seeks to identify those at greatest risk of ill-effects from specific environmental exposures to toxins and other external agents on the basis of their genomic composition and activity.

Perhaps more significant at this stage is pharmacogenomics, common genetic variations that affect the components of drug metabolic pathways in individuals and govern drug responses. This can limit or prevent the efficacy of standard doses of some drugs in 'fast metabolisers', but it can be even more important to identify 'slow metabolisers', in whom a standard dose could have serious adverse effects, via pharmacogenetic testing. For example, over 10% of the UK population has genetic variants in the thiopurine methyltransferase (TPMT) gene that reduce their ability to metabolise immunosuppressant thiopurine drugs, putting them at risk of dangerous toxicity – unless revealed by testing.

Genetic testing can also be of value with respect to infectious disease risks, to identify individuals with genetic variants that make them particularly susceptible to specific infectious disease agents, or to severe forms of disease related to specific infections. If patients are known to be at risk, they can be prioritised for prompt and aggressive treatment in the event of a possible infection.

In the future, genomic data may also be relevant to health-related behavioural modification. Genetics certainly influences individual propensity to obesity, and responses to dietary intake. Nutrigenomics, the study of genetic contributions to individual variation in nutritional requirements and responses, could eventually help identify those at increased risk of developing conditions such as metabolic syndrome or diabetes, as well as tailor nutritional recommendations to minimise disease risk or progression.

A bigger challenge lies in making sense of genetic influences on mental health and behaviour, including propensity towards harmful forms of addiction. Large-scale genomic research into psychiatric disorders is revealing shared origins that could have a major effect on how different disorders are classified and treated. Eventually such data could also have some predictive utility, although use for this purpose could be controversial.

## The future of individualised risk prediction

Genomic data undoubtedly have an important role to play in disease risk prediction and personalised medicine in the coming years. The better we understand genomics and the underlying biological processes that cause disease, the more opportunity there is to predict, prevent and treat it most effectively, as well to develop new and more tailored treatments. However, for most diseases genomic data remains just one piece in the puzzle.

The scope for truly individualised prediction and prevention of common disease is likely to be some way off, and will probably need to combine not only basic genetic and lifestyle information, but also data from a much wider range of sources, whether fitness trackers or other forms of medical wearables, in risk prediction algorithms. Just as new technologies and applications are being advanced as an opportunity for more personalised health promotion via behavioural modification and psychological support, others such as implantable or portable biosensors and scanners (perhaps via smartphones) could contribute a multiplicity of health-related data that, properly streamed and analysed via predictive algorithms, could help provide refined risk estimates – whether static (at a given point in time) or potentially even continuous. Indeed, in examining data from biomarkers that may be used to measure multiple bodily processes, from gene expression to metabolism or even the presence or absence of specific microorganisms in different parts of the body infections, it may be not so much absolute levels as patterns of change in an individual that are most predictive of changes in risk, or that can identify pre-disease states.

Eventually, we may progress to a state where risk prediction forms just one part of a continuous lifelong process of personal monitoring and health maintenance – but there is still a long way to go before we get there.

## References


Chapman S.J., Hill A.V.S. (2012). *Human genetic susceptibility to infectious disease.* Nature Reviews Genetics 13: 175-188.

Cross-Disorder Group of the Psychiatric Genomics Consortium (2013). *Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs.* Nature Genetics 45: 984-994.

Joly Y., Burton H., Knoppers B.M. et al. (2011). *Life insurance: genomic stratification and risk classification.* European Journal of Human Genetics 22(5): 575-579.

Lennard L. (2014). *Implementation of TPMT testing.* British Journal of Clinical Pharmacology. 77(4): 704-714.

National Institute for Health and Care Excellence (2008). *Familial hypercholesterolaemia: identification and management.* London: NICE.

National Institute for Health and Care Excellence (2013). *Familial breast cancer: classification, care and managing breast cancer and related risks in people with a family history of breast cancer.* London: NICE.

Pashayan N., Duffy S.W., Chowdhury S. et al. (2011). *Polygenic susceptibility to prostate and breast cancer: implications for personalised screening.* British Journal of Cancer 104(10): 1656-1063.

Pirmohamed M. (2012). *Genetics and the potential for predictive tests in adverse drug reactions.* Chemical Immunology and Allergy 97: 18-31.

Zhang R., Chu M., Zhao Y. et al. (2014). *A genome-wide gene-environment interaction analysis for tobacco smoke and lung cancer susceptibility.* Carcinogenesis. 35(7): 1528-35.


## Biography

**Philippa Brice** is an External Affairs Director at PHG Foundation. She trained in Natural Sciences (Pathology) at Christ's College, Cambridge before completing her doctorate at the MRC Laboratory for Molecular Biology and moving to work in the pharmaceutical industry. She is a Senior Member of Hughes Hall.

# 8. Big data in action: wearables

Matthew Edwards, Senior Consultant, Willis Towers Watson

The commonly accepted characteristics of what constitutes 'Big Data' – rather than just a big dataset – are the four Vs': volume, velocity, variety and the somewhat forced fourth alliterative attribute, veracity. Volume refers to the quantity of data – for instance, Facebook is thought to deal with of the order of 3 billion 'likes' or comments every day. Velocity refers to the frequency of updating – for instance, daily feeds rather than monthly or yearly updates – and hence the required velocity of analysis if data use is to be optimised. Variety relates to the nature of the data: structured, unstructured; numbers, text, pictures, etc. Veracity is a question not just of data validity per se, but the extent to which there is a real 'signal' underlying an undoubtedly large amount of noise.

It is clear from these criteria that the mass of data being gathered from wearables qualifies as big data. This is true even if we consider just wrist-worn activity-tracking devices such as Fitbits and smart watches, but the definition could also extend to smartphones with activity-tracking apps, adding vastly more devices to the data-generating pool. The volumes are enormous, with of the order of hundreds of millions of the leading wearables in circulation; velocity is provided by daily (if not more frequent) uploading of the wearer's activity; variety comes from the range of fields, typically relating to steps walked/run, sleeping patterns and circulatory information; on the other hand, veracity remains something of a challenge for wearables, given issues around accuracy (although wearable accuracy continues to improve) and the problem, in the context of possible insurance uses, of ensuring the wearer is indeed the 'named' wearer.

Wearables present a fascinating opportunity for insurers, and related sectors such as the private health sector.
In the first instance, wearables could be regarded as the life insurance equivalent of the vehicle telematics and usage-based insurance that has become relatively common in motor insurance, providing enormous amounts of information about the insured's driving habits – and also providing an immediate form of segmentation (a policyholder who accepts a telematics device in their car will likely be a materially different risk type from the refusenik). The nature of the insured event, however, makes the two applications very different – information on a driver's top speed compared with speed limits may be predictive in motor accident analyses, but how fast a 'wearable wearer' sometimes sprints is unlikely to be similarly predictive in a morbidity or mortality analysis.

At the moment, there are several areas where wearables present opportunities for insurers, and several insurers in the UK and the US are moving in some of these. The primary areas are risk assessment and in-force management.

The information provided by wearables can clearly be of use in fine-tuning morbidity/mortality models, and hence personalising rates in a relatively non-invasive way – providing wearable data is more 'consumer friendly' than providing fluid samples. There are, equally clearly, practical issues around tying data to the right individual, and ensuring that low-priced policies are, if possible, tied to continuation of a healthy lifestyle.

The second area of in-force management has a similar aim, of seeking to provide policyholders with a health discount to their premiums, but avoids some of the practical issues associated with the above risk assessment aspect. It can also assist with long-standing policyholders – improving the health of an insurer's portfolio or an employer's workforce (from the perspective of group life and disability insurance costs, quite apart from productivity). These are good things both from the corporate perspective, be it the insurer or the employer, but also that of the individuals in question.

Examples of firms known to be operating along these lines include AXA, where two of its US units (AXA Equitable Life Insurance Co. and MONY Life Insurance Co. of America) boast a Wellness Incentive Benefit Endorsement. Under this, policyholders can receive payments on completing specified health activities, such as regular exercise while wearing an approved fitness-tracking device. The US insurer John Hancock has a similar benefit, termed the Healthy Engagement Benefit: policyholders can reduce their premiums from scoring sufficient points, for instance via exercise as monitored according to an approved wearable, or achieving and maintaining a healthy body mass index.

In the UK, the relatively new insurer Vitality is regarded as the leader in this area, with its 'Vitality Optimiser' which offers a range of benefits to policyholders who choose this route. Interestingly, Vitality has been the first UK insurer to partner with Apple Watch.

So far, much of the premium consideration relating to initial discounts on wearable use is thought to be fairly broad brush, being designed more as an incentivising mechanism to attract the desired type of policyholder. Marketing is itself a perfectly reasonable justification for 'using' wearable device data, and it can help develop, along with other wellness initiatives, important affinity relationships to position the insurer as a preventer of risk.

Where premium discounts are considered more scientifically, the calculation of appropriate reductions relating to (for instance) daily steps walked requires a decent understanding of the relationship between those metrics and the claim outcome of interest, whether mortality or health-related claims. Given the novelty of electronic wearables and the fact that they tend to be used by fitter, younger lives, even a large insurer would need to wait many years before achieving the critical mass of mortality data that would enable interesting multi-factor analyses. However, substantial research has been done on the health effects of exercise, in particular steps, using the 'old fashioned' routes of either pedometers or self-declared exercise levels. This is further illustrated in Figure 7.

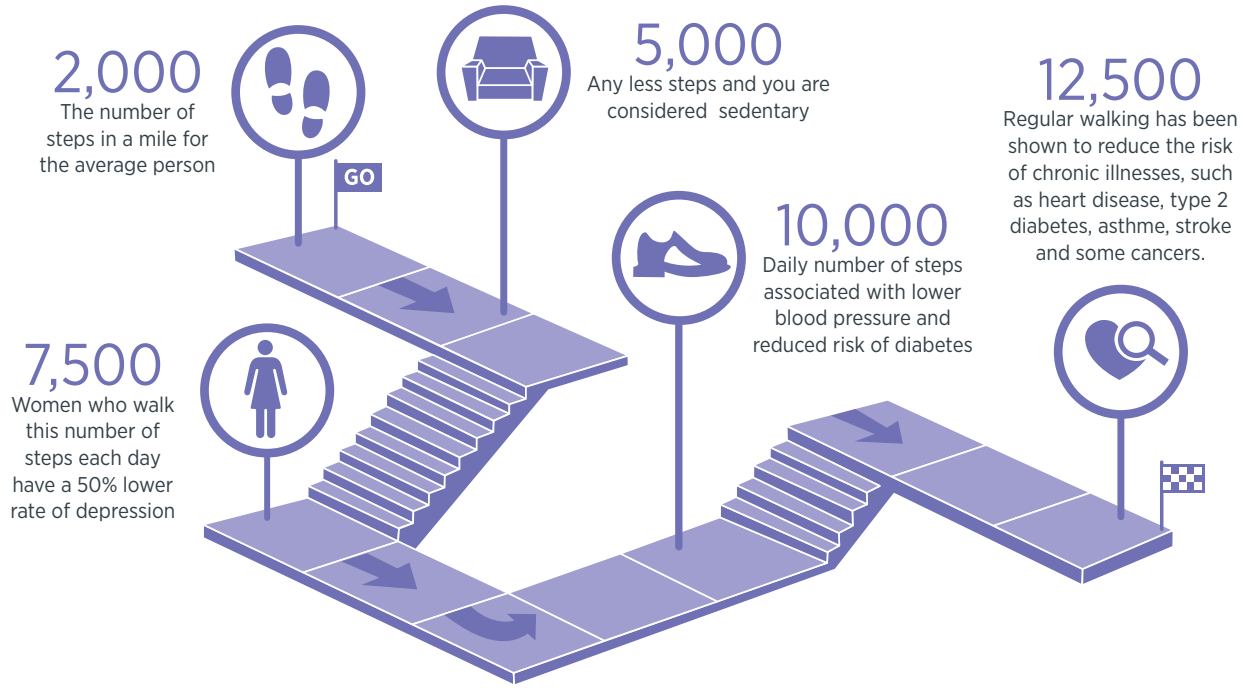
By way of examples (although these relate primarily to the health benefits for the currently sedentary, rather than the many current users of wearables who are already fairly fit):

- Each 2,000 step per day increment is associated with a 10% reduction in probability of a cardiovascular event. (Yates, Haffner et al., 2014)
- Increasing baseline daily steps from sedentary to 10,000 / day associated with 46% reduction in mortality (Objectively measured daily steps and subsequent long-term all-cause mortality, (Dwyer, Pezic et al. 2015)
- The relative risk of death (all-cause mortality) for a 65-year old U.S male with a vigorous (4 miles per hour) walking speed compared with a slow (2-2.5 miles per hour) pace was 0.9 (Olshansky, 2016)

The most recent news in the field of predictive health analytics based on wearables comes not from insurers but from the North American health sector.

For instance, the firm 'LifeQ', based in Georgia, has developed a new type of optical sensor that can be integrated into wearable devices to monitor various physiological measures. The data captured by this sensor is then used as inputs into already-designed models to provide 'real time' information on the body's health and performance.

**Figure 7: Stepping to health**




Source: Catrine Tudor-Locke, Walking Behaviour Laboratory. NHS – Walking for Health

(With thanks to Vitality Life for reproduction permission)

The Canadian firm Vivametrica, based in Calgary, was established to analyse data from wearable sensor devices in order to improve health and wellness. The Public Health Agency of Canada has found that physically active employees take on average 27 per cent fewer sick days and 14 to 25 per cent fewer disability-due-to-injury days than inactive employees: the potential for corporate and individual benefits from wearable usage and analysis seems clear.

Given their ability to accurately assess chronic disease risk and mortality, Vivametrica has recently shifted their focus to application of their proprietary health analytics for underwriting and engagement of insurance customers.

The firm has constructed a device-agnostic data system to acquire, standardise and analyse data from wearable technology. Their algorithms provide the users with insights into their health, doing so with emphasis also on data security and privacy (commercial usage of their data is done only in an anonymous and aggregated fashion with the individual's consent). Vivametrica uses various 'big data' analytic techniques including traditional epidemiological methods, combined with machine learning. Their proprietary population-based database includes hundreds of thousands of individuals, and billions of behavioural and clinical data points.

The infancy of electronic wearables means that there is little published research available. It is clearly an emerging 'Big Data' field of great interest to insurers and healthcare providers, as well as to medical researchers – and of course to desk-bound office workers looking to improve their health.

## References


Yates, T., Haffner et al. (2014). *Association between change in daily ambulatory activity and cardiovascular events in people with impaired glucose tolerance (NAVIGATOR trial): a cohort analysis.* The Lancet 383 (No. 9922) March 22: 1059-1066.

Dwyer, T., Pezic, A. et al. (2015). *Objectively measured daily steps and subsequent long-term all-cause mortality: the Tasped prospective cohort study.* PLOS ONE 10(11): e0141274.

Olshansky, J. (2016). International Mortality and Longevity Symposium [2016]. Institute and Faculty of Actuaries, 2016.


## Biography

**Matthew Edwards** is the Head of Mortality and Longevity in Willis Towers Watson's life insurance practice. He has a particular interest in disease-based modelling, longer-term driver-based models of longevity, and using the views of medical experts to enhance our understanding of likely mortality trends.

# 9. The ethical challenges of biomedical data

Professor Luciano Floridi, Oxford Internet Institute, University of Oxford

In biomedical research, the analysis of large datasets (big data) has become a major driver of innovation and success. However, the use of biomedical big data (BBD) also raises serious ethical problems, which may threaten the huge opportunities it offers. The risk is that of a double bottleneck: ethical mistakes or misunderstandings may lead to distorted legislation, which may cripple the usability of BBD in medical research, health care, and industry, as evidenced by a recent statement issued by the Wellcome Trust on "The impact of the draft European Data Protection Regulation and proposed amendments from the rapporteur of the LIBE committee on scientific research". As a consequence, there is a widely acknowledged need for a European framework for the ethical use of big data in biomedical research. Three main research objectives should be pursued:

1. to formulate a blueprint of the ethical aspects, requirements and desiderata underpinning the project for a European framework for the ethical use of big data in biomedical research;
2. to strengthen and coordinate multidisciplinary research in the area of ethical and relevant socio-legal aspects of BBD; and
3. to consolidate world-leading expertise in the ethics of BBD that will enable and support research on large health-related datasets (e.g. at the Li Ka Shing Centre for Health Information and Discovery) and will contribute to the goals of the Strategy for the UK Life Sciences, which aims to improve research outcomes and the attractiveness of the UK as a centre for global research excellence.

On November 28-29 2012, the University of Oxford hosted the Oxford-Stanford Conference on Big Data: Challenges and Opportunities for Human Health. Supported by the Li Ka Shing Foundation. The Conference concluded that:

*"we are poised for a revolution in how we understand disease and treat patients in the 21st century."* (University of Oxford and Stanford University, 2012)

Philip Pizzo (The Carl and Elizabeth Naumann Professsor of Pediatrics and Microbiology & Immunology and Dean of the Stanford University School of Medicine) and John Bell (Regius Professor of Medicine) remarked that:

*"In many areas of science, government and business, analysing large amounts of information – Big Data – has become a major driver of innovation and success. In biomedical research new technologies and collaborative approaches mean we are facing our own data revolution with the potential reward of major improvements to human health and healthcare."* (University of Oxford and Stanford University, 2012, p. 26.)

According to Alastair Buchan (Professor of Stroke Medicine and Head of the Medical Sciences Division), the goal of the meeting was:

*"making a step change for medicine in the 21st century, in our ability to interpret eloquent signals from these very large datasets."* (University of Oxford and Stanford University, 2012, p. 27.)

Epidemiology (Salathé et al., 2012), infectious diseases (Hay et al., 2013), and genomics and genetics (Watson et al., 2010), are already deeply affected by Biomedical Big Data (BBD) (Floridi, 2012). Unfortunately, as John Bell acknowledged:

*"[BBD in these three areas represents a] huge opportunity for major historical advances but it will only come if you can analyse the data in some sensible way."* (Bell, 2012).

The Conference identified ethics as one of the most significant challenges for such a *"sensible way"*.

The use of BBD raises several ethical problems (Safran et al., 2007), both sensitive and complicated. They may be clustered under six headings.

**1) Deficit.**

There is an acknowledged lack of public awareness of the benefits, risks, and challenges associated with BBD; of transparency of use of BBD for purposes other than direct patient care and public health; of a clear and shared taxonomy for secondary uses (including non-clinical ones)

of personal health information and electronic health records in order to clarify ethical issues; and of policies, practices, as well as safeguards that adequately address secondary uses of BBD.

**2) Consent.**

It is unclear which kind of informed consent (broad, specific, dynamic, blanket…) may be preferable when it comes to how BBD may be used for specific purposes or re-purposed, in compliance with legislation; and more work is required about patient choice options involving explicit authorisation for use of their health data (opting in/opting out) to mitigate privacy issues.

**3) Privacy.**

Apart from well-known concerns, BBD give rise to two new problems. There is a risk of re-identification of patients and providers through data-mining, data-linking, data-merging and re-using of large datasets. And there is a risk about "group privacy", when the identification of types of individuals, independently of the de-identification of each of them, may lead to serious ethical issues, from group discrimination (e.g. ageism, ethnicism, sexism) to group-targeted forms of violence, especially in areas of the world politically unstable or undemocratic.

**4) Coverage.**

Data uses may not be covered by ethical regulations (especially when data are obtained via coerced or compelled consent) causing the erosion of public trust and confidence.

**5) Balance.**

In light of serious public health threats (e.g. avian flu) and bio-risks (cf. anti-terrorist biosurveillance), how the duty to protect and enhance the public good may be reconciled with the rights of individuals (public health versus individual privacy); then there is a medical version of the notorious digital divide about the right to participate in and benefit from BBD-based research.

**6) Management.**

Who has the right to access, use, audit, control (e.g. constrain the use and repurposing of), release (e.g. in the public domain), and own (including Intellectual Property Rights of derivate products) what health data, for what purposes, and at what stage in the data life cycles, including metadata, and further data that are generated by primary data?

A further difficulty is that the previous ethical problems are multidimensional, as they need to be mapped across two axes (Davis & Patterson, 2012; Groves et al., 2013).

- One axis is represented by the interactions between medical research, health care practice and delivery, and the commercialisation of health data and use of health data for business and proprietary purposes. For example, when is it ethically right or even obligatory to make some large datasets available across a scientific laboratory, a hospital, and a pharmaceutical company?
- The other axis is represented by pre-existing normative guidelines in medical, business, and research ethics, professional codes of conduct, and accepted best practices within the corresponding communities of users. For example, when is it ethically right or even obligatory to inform patients about some BBD results that may affect the health conditions of their progeny?

Given such a complex scenario, the use of BBD is threatened by a double bottleneck: ethical mistakes or misunderstandings may lead to distorted legislation, which may cripple the usability of BBD in medical research, health care, and industry. This *"nested dolls problem"*, in which ethics is the outer layer that constrains legislation, which in turn is the outer layer that constrains medical research, means that ethical issues become *"The metaphorical 'Thermopylae' of many biomedical research projects. […] limited experience and understanding of many of the relevant issues […] leads to serious misunderstandings and delays, particularly when norms are applied that are simply not suited to the real nature of biobanks."* (Khoury, 2010, p. 89).

A recent Statement issued by the Wellcome Trust on *"The impact of the draft European Data Protection Regulation and proposed amendments from the rapporteur of the LIBE committee on scientific research"* (Wellcome Trust, 2013) is indicative of the risks involved and of the damaging consequences that such misunderstandings may have.

The Oxford-Stanford Conference Report concluded that,

*"to seize the opportunities offered by Big Data in medical research and health care [institutions need to] work with legal, ethical and political interests to ensure that Big Data projects are incorporated into research and health care in a manner that benefits society."* (University of Oxford and Stanford University, 2012, , p.10.).

The formulation of the ethical requirements and the need to promote the beneficial uses of BBD in medical research, health care, and industry must strike the right balance between the moral obligation to pursue therapeutic ends and the uncertainty about the consequences of the technical means, and hence between the researchers' duty to exercise benevolence and the patients' right to see caution applied in the proper use of BBD. As indicated by the Nuffield Council of Bioethics in the context of the ethical framework for neurotechnologies (Nuffield Council of Bioethics, 2013) in articulating the implications of the principles of beneficence and caution, particular attention needs to be paid to five interests: potential safety risks, unintended impacts on privacy, the promotion of autonomy, public interest in equity, and public understanding of trust.

Clearly, the development of the framework for the ethical use of big data in biomedical research is a difficult task. It may be an even more complex when considering the challenges that non-medical organisations face in accessing big data. But it is one that needs to be undertaken as soon as possible (for some initial steps see Mittelstadt and Floridi, 2016a and 2016b).

## References


University of Oxford; Stanford University (2012). *Report: The Oxford-Stanford Conference on Big Data: Challenges and Opportunities for Human Health*. 2012.

Salathé, M., et al. (2012). *Digital Epidemiology.* PLoS computational biology, 2012. 8(7)

Hay, S. I., George, D. B., Moyes, C. L. and Brownstein, J. S. (2013). *Big Data opportunities for global infectious disease surveillance.* PLoS medicine, 10(4).

Watson, R. W. G., Kay, E. W. and Smith, D. (2010). *Integrating Biobanks: addressing the practical and ethical issues to deliver a valuable tool for cancer research.* Nature Reviews Cancer 10(9): 646-651.

Mathaiyan, J., Chandrasekaran, A. and Davis, S. (2013). *Ethics of genomic research.* Perspectives in Clinical Research 4(1): 100.

Floridi, L. (2012). *Big Data and their epistemological challenge.* Philosophy & Technology 25(4): 435-437.

Bell, J. (2012). *Big Data - challenges & opportunities for human health.* YouTube interview, 2012.

Howe, D., et al. (2008). *Big Data: the future of biocuration.* Nature 455(7209): 47-50.

Safran, C., et al. (2007). *Toward a national framework for the secondary use of health data: an American Medical Informatics Association White Paper.* Journal of the American Medical Informatics Association 14(1): 1-9.

Davis, K. and Patterson, D. (2012). *Ethics of Big Data.* Farnham: O'Reilly.

Groves, P., Kayyali, B., Knott, D. and Van Kuiken, S. (2013). *The 'Big Data' revolution in healthcare.* McKinsey Quarterly.

Khoury, M. J. (2010). *Human genome epidemiology: building the evidence for using genetic information to improve health and prevent disease.* 2nd ed. New York: Oxford University Press.

Wellcome Trust (2013). *The impact of the Draft European Data Protection Regulation and Proposed Amendments from the Rapporteur of the Libe Committee on Scientific Research".* Available online.

Nuffield Council on Bioethics (2013). *Novel neurotechnologies: intervening in the brain.* June 2013. Available online.

Mittelstadt, B. D. and Floridi, L. (2016a). *The ethics of Big Data: current and foreseeable issues in biomedical contexts.* Science and Engineering Ethics22(2): 303-341.

Mittelstadt, B. D. and Floridi, L. (eds) (2016b). *The ethics of Biomedical Big Data.* New York: Springer, 2016.


## Biography

**Professor Luciano Floridi** is Director of Research and Professor of Philosophy and Ethics of Information at Oxford Internet Institute (OII). His long-term project is a tetralogy on the foundations of the philosophy of information. Before joining the OII, he was Chairman of the European Commission's expert group "Concepts Engineering", on the impact of information and communication technologies on the digital transformations occurring in the European society.

# 10. Recent developments and events

## News from the IFoA

Working parties update:

The Institute and Faculty of Actuaries is active in supporting volunteer-led groups (also known as working parties) to conduct research on timely topics of wide interest to the practice area and profession. Working parties currently conducting research in the field of big data are listed below.

### Impact of Wearables and the Internet of Things Working Party

Earlier this year a working party was set up to look at the emergence of wearable technology and the internet of things, and the current and potential use within the health and care area.

This is a rapidly developing area of technology, with new devices and advances coming on to the market regularly.

We have decided to focus on three areas for this research;

**1) Understanding the stakeholders**

By considering the interests of the various stakeholders (consumers, employers, distributors, (re)insurers, and manufacturers), we will investigate the areas of conflict and commonality between them, with an aim to understand the potential challenges and opportunities in linking this technology with health insurance.

**2) Current & future state of technology**

We will research the various technologies that are currently available which have a potential to impact health and wellbeing, and also what the next generation of these may look like.

This will include looking at;

- What can be measured
- What data is captured and how
- How accurate and reliable the data is
- How end users engage with this technology
- Behavioural impact in the short and long term.

**3) Practical applications in health**

Finally we will look at the practical use and impact of this technology in the life and health insurance market. What areas of the insurance cycle could this technology and the data it provides be used in: underwriting, pricing, proposition development, rehabilitation and claims management or capital/risk management? And what challenges companies may face using this data in a practical way?

The first two phases of research will run concurrently with an aim to produce an article for The Actuary. It is envisioned that the working party will present the full research at the Protection Health and Care Conference in 2018.

If you are interested in joining this working party or your feel you can contribute to this research please contact us. Details can be found on the IFoA webiste: **http://bit.ly/2gSZZOm**

### Modelling, Analytics and Insights from Data (MAID) Working Party

Big data often means the whole plethora of new techniques for investigating what this data can tell its users. The recently formed MAID cross practice working party focuses on the latest and emerging thinking associated with modern mathematical tools and techniques and explores how the actuary may utilise these techniques to remain practical and add value to the businesses it serves.

This working party now has some 65 volunteers organised into four work streams. Our terminology is that the data science universe is the best wording to describe the coming together of computer science, mathematics and operational research. Indeed these previously separate academic disciplines are starting to offer combined programmes – such as machine learning, or data visualisation. It would be brave to predict what will emerge from this – but it will revolutionise the work of actuaries.

The four work streams cover:

- **Work stream 1 (Research)** has survey members of the IFoA and is analysing findings around how communication of data

science issues can be improved, CPD enhanced and generally whether the profession sees an opportunity or a threat. Its next task is a literature review.

- **Work stream 2 (New approaches to existing actuarial problems)** is looking into four potential case studies and aims to publish a summary early in 2017.

  These case studies are:

  - **Marine Hull** - Learning how machine learning can better enhance the accuracy of Pricing models and which features impact claims from different ships
  - **Exposure Management** - seeing how predicting missing fields 'Year Built' and 'Building Stories' has an impact on calculating Estimated Loss in Catastrophe models
  - **Life and Mortality** - linking external data sources such as Dow Jones, Consumer Confidence and Mood indices with US death records
  - **Strategic/Tactical Asset Allocation and Asset & Liability Management / Hedging**

- **Work stream 3 (New opportunities for actuaries)** is looking at opportunities beyond current work areas such as new aspects of life or general insurance, banking or even something as generic as customer database analysis in a wide sense. It is also looking into computer and technology developments.
- **Work stream 4 (Implications for our profession)** has presented an update to the IFoA's Management Board in August and is developing a discussion paper on what strategy the IFoA might adopt to the opportunity. This would cover exams, CPD, life long learning and accreditation, PR, approach to consultations, internal regulation and professional standards, recruitment and member engagement. It is also mapping organisations we as a profession should work with and connect to.

To find our more about the MAID working party please visit the IFoA website here: **http://bit.ly/2h6BhKT**

### E-cigarettes Working Party

A working party was set up in July 2016 to investigate the impact of e-cigarettes along with other related reduced risk tobacco products.

E-cigarette usage has dramatically increased in recent years from 0.7 million in 2012 to 2.8 million UK (source: ASH estimate, 2016). The overall impact on health implications is uncertain:

- the relative impact of these products compared to traditional cigarettes seems to be at around 90% to 95% less in terms of disease, however;
- there is an unknown impact around behavioural changes made by smokers;
- and public perception is confused!

There is an ongoing debate on the health impact for individuals which could potentially lead to a major contribution towards preventing premature death, disease and social inequalities in health that smoking currently causes in the UK.

The working party plan is to present at forthcoming industry events including the Health and Care conference in May 2017.

Further details can be found on the IFoA website here: **http://bit.ly/2gjOG0U**

### Who, when and why? Mortality datasets provide a wealth of information for actuaries

The use of mortality data lies at the core of much of the research carried out by the actuarial profession. In the area of health, longevity and mortality many readers will be aware of commonly used datasets such as those from the Office of National Statistics (ONS) or WHO. There is, however, a wealth of mortality data available, much of it free to access and available online. The Institute and Faculty of Actuaries has compiled a directory of datasets which cover the UK and Ireland as well as those that give an overview of European and world data. The directory not only lists the datasets and provides links to each but also provides some detail on data points of interest to actuaries and the timeframe over which the data was collected.

It is hoped that greater access to and awareness of these datasets will enable more accurate modelling, allowing actuaries to make informed decisions regarding longevity and mortality in relation to life assurance, pensions and long term care products.

To access this free resource please visit: **http://bit.ly/2gle7h4**

## Upcoming events

### Save the date: 2017 IFoA Spring Lecture on antimicrobial resistance

Dame Sally Davies, Chief Medical Officer for England

27 April 2017, 17:30 – 20:00, London

Antimicrobial resistance (AMR) has increased alarmingly, accelerated by the overuse of antibiotics in many countries for medical and also agricultural purposes. In the IFoA 2017 Spring Lecture, Dame Sally Davies will explore why AMR has developed to such an extent that it is now a threat to modern medicine.

Further information will be published on the IFoA website in due course: **http://bit.ly/gpdpa**

## Call for speakers: Protection, Health and Care 2017 Conference

The Protection, Health and Care Conference is an annual conference aimed at all insurance professionals with a passion for harnessing insurance risk in their organisations.

Following a successful Protection, Health and Care 2016, we are now working on the programme for next year and we are looking for a wide variety of topical workshop sessions.

If you would like to speak at next year's conference please submit your proposal on the IFoA website here: **bit.ly/2grv4VG**

Closing date: 12 January 2017.

## Mortality and Longevity Seminar

**22 June 2017, London**

Following on from a successful and well-attended workshop in 2016, this year's seminar will appeal to Pensions, Life and Health and Care actuaries eager to learn about the latest developments and current 'hot topics' in mortality and longevity.

If you're interested in presenting, please submit your proposal on the IFoA website here: **bit.ly/2grv4VG**

Closing date: 11 January 2017.


Institute
and Faculty
of Actuaries

**London**

7th Floor · Holborn Gate · 326-330 High Holborn · London · WC1V 7PP
**Tel:** +44 (0) 20 7632 2100 · **Fax:** +44 (0) 20 7632 2111

**Edinburgh**

Level 2 · Exchange Crescent · 7 Conference Square · Edinburgh · EH3 8RA
**Tel:** +44 (0) 131 240 1300 · **Fax:** +44 (0) 131 240 1313

**Oxford**

1st Floor · Park Central · 40/41 Park End Street · Oxford · OX1 1JD
**Tel:** +44 (0) 1865 268 200 · **Fax:** +44 (0) 1865 268 211

**Beijing**

6/F · Tower 2 · Prosper Centre · 5 Guanghua Road · Chaoyang District · Beijing China 100020
**Tel:** +86 (10) 8573 1522

**Hong Kong**

1803 Tower One · Lippo Centre · 89 Queensway · Hong Kong
**Tel:** +852 2147 9418

**Singapore**

163 Tras Street · #07-05 Lian Huat Building · Singapore 079024
**Tel:** +65 6717 2955

www.actuaries.org.uk