



**Mario Cortina Borja** is chair of the *Significance* editorial board, and professor of biostatistics in the Population Policy and Practice Teaching and Research Department at the Great Ormond Street Institute of Child Health, University College London.

that we have predictor variables measured on each observation (e.g., age, health status, biomarkers). We would like to carry out a regression analysis to assess the effects of each predictor; if one of the predictors is one of two drugs, we would like to use regression to compare the two drugs while adjusting for confounder factors.

Cox's model is based on a quantity known as the *hazard function*  $h(t | x)$ . This is the probability that an individual with predictors  $x$  will die at time  $t$ , given that the individual is alive just before  $t$ . Cox's regression model starts with an assumption of *proportional hazards*:  $h(t | x) = h_0(t)\exp(x\beta)$ . This says that the hazard for an individual with predictors  $x$  is the product of a baseline hazard  $h_0(t)$  (corresponding to  $x = 0$ ) and a factor that depends on  $x$  and the regression parameters  $\beta$ .

To estimate  $\beta$ , Cox invented a *partial likelihood*: it uses the same “product in time” construction as in the Kaplan–Meier method, computing the probability under the model of surviving past each event time  $t_k$  as a product of conditional probabilities. This is a function of the unknown regression parameters  $\beta$ , and this function is maximised (with a numerical optimiser) to yield the partial maximum likelihood estimates  $\hat{\beta}$ . Furthermore, this elegant construction does not require specification of the baseline hazard  $h_0(t)$ , making the method flexible and robust.

With Cox's proportional hazards model we obtain all of the things that we enjoy in regression analysis, such as parameter estimates, standard errors and confidence intervals. But there is much more: it provides a natural way of modelling *time-dependent covariates*, where a patient characteristic (such as blood pressure) is measured not once at the beginning of a study but many times over the course of a study. And it can deal with other forms of censoring such as left and interval censoring. The method has also been generalised to handle high-dimensional data, incorporating sparsity penalties such as the lasso.<sup>37,38</sup> And there are even deep learning versions of Cox's method.<sup>39</sup>

The impact of this method on many fields, especially medical sciences, has been enormous; it surely will continue to grow in importance in the future. ■

#### Disclosure statement

The author declares no competing interests.

## On Sir David Cox's publications

**Mario Cortina Borja, Julian Stander, and Giovanni Sebastiani** analyse some of the tens of thousands of citations for the hundreds of papers produced by David Cox in a publishing career that spanned more than 70 years

**S**ir David Cox's publishing career was remarkable for the extent and influence of its contributions, as well as for its duration; his works have been published continually from 1947 to 2021. David's website at Nuffield College, Oxford, lists 385 publications ([bit.ly/3M9eIB4](http://bit.ly/3M9eIB4)) and we shall refer to this collection as N385. These numbers are impressive on their own, but we should also remember and celebrate David's unflinching kindness to generations of statisticians and other scientists, of which there are many accounts in this issue of *Significance*. Here we examine in more detail the extent and influence of David's work by focusing on trends in the number of references or mentions of his works in scientific publications. We will pay particular attention to his paper on “Regression models and life-tables”,<sup>14</sup> which we will refer to as Cox (1972).

### Data

The publications in N385 comprise 302 peer-reviewed articles, 23 books, and 60 entries classified as chapters in edited books, entries in encyclopaedias, proceedings, conference proceedings, reports, or lectures. (Publications are numbered from 1 to 384 on the website, but there are two entries with the number 191.) This publication list is incomplete as it does not include, for example, David's many contributions to the discussions of Royal Statistical Society read papers. His peer-reviewed output appeared in 104 different journals. There were 56 papers in the *Journal of the Royal Statistical Society* (*JRSS*; 35, 13, 7, and 1 in *Series B, A, C, and D*, respectively), 60 in *Biometrika*, and 24 in International Statistical Institute publications.

We also analysed data from two sources: Web of Science (WoS), which claims to be “the world's most trusted publisher-independent global citation database”, and PubMed (PM), which “comprises more than 33 million

citations for biomedical literature from MEDLINE, life science journals, and online books”. These data were downloaded between 5 and 21 February 2022.

For each year after publication, we obtained information on *citations* from WoS, where the number of citations of one of David's outputs is the number of WoS papers listing the output as a reference. Similarly, PM yielded data on *mentions*, that is, the number of PM papers containing text referring to the output. A paper that contributes a mention does not necessarily contribute a citation. Below we consider mentions of Cox (1972), identified by searching on {“Cox regression” OR “Cox's regression” OR “Cox proportional hazard\*” OR “Cox's proportional hazard\*”}.

### Trends in citations

WoS returned 74,406 citations for papers in N385. Figure 1 shows the cumulative number of citations for David's eight most cited papers. The two most cited papers, Cox (1972) and the Box–Cox (1964) paper on the analysis of transformations,<sup>12</sup> were published in *JRSS Series B*. These are followed by two papers with several co-authors in the *British Journal of Cancer* on the design and analysis of randomised clinical trials,<sup>40,41</sup> and David's (1975) *Biometrika* paper in which partial likelihood is defined.<sup>15</sup> This list concludes with three more *Series B* articles: a pivotal paper defining logistic regression for binary sequences in 1958;<sup>5</sup> a 1987 paper with Nancy Reid on parameter orthogonality and approximate conditional inference;<sup>18</sup> and a 1968 work with E. Joyce Snell proposing a general definition of residuals for linear models.<sup>42</sup> The sharp increase in the last five years of citations of David's (1958) analysis of binary sequences paper is driven by references made in areas including bioinformatics, machine learning, remote sensing, and neuroscience where logistic regression is mostly used as a classification method. ▶



**Julian Stander** is an associate professor of mathematics and statistics at the School of Engineering, Computing and Mathematics, University of Plymouth.



**Giovanni Sebastiani** is senior researcher at the Istituto per le Applicazioni del Calcolo "M. Picone" of the Italian National Research Council.

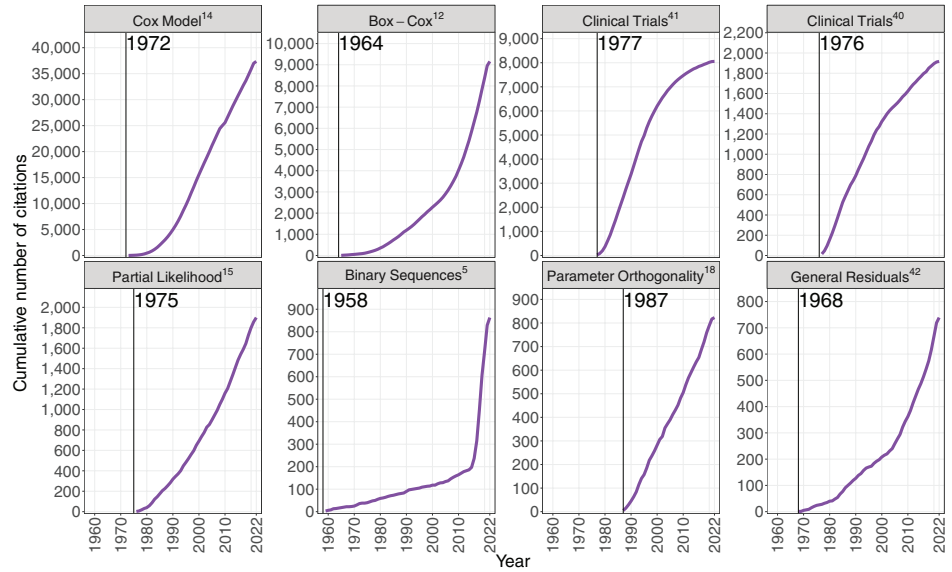
### ► Doubling time

We now concentrate on mentions of Cox (1972), which first appeared in PM in 1978. Remarkably, out of the 30.5 million papers contained in PM between 1978 and 2021, over 150,000, or almost 1 in 200 (0.5%), mentioned Cox (1972). Indeed, the percentage of PM papers mentioning Cox (1972) steadily increased from 0.18% of papers published in 2001, to 0.52% for 2011 and 1.15% for 2021.

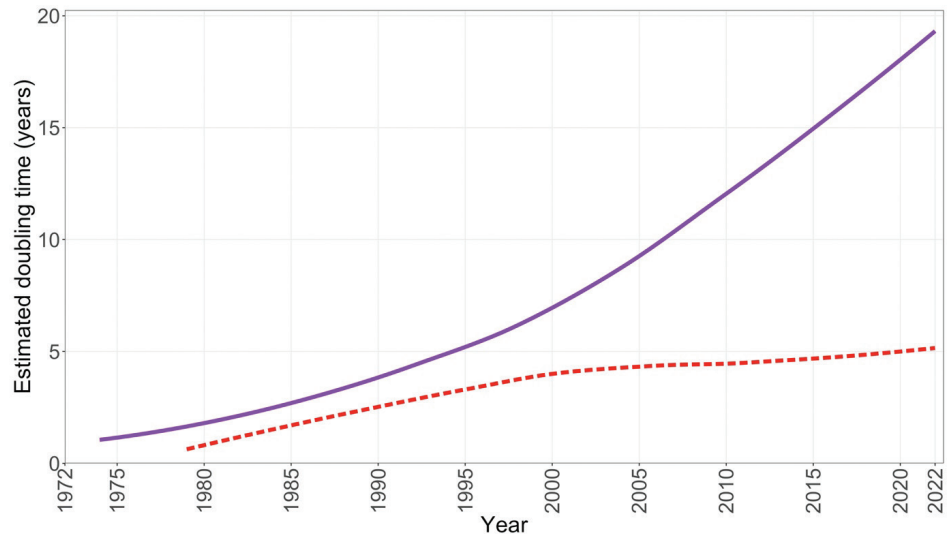
One way to understand increases in citations and mentions is through the concept of doubling time, which is the shortest time in years needed for the number of items to double. Because we cannot see into the future, we estimate doubling time by working backwards to find when the number of mentions was at half its current value (as discussed in a previous article: [significancemagazine.com/657](http://significancemagazine.com/657)). The number of years that we need to go back is therefore an estimate of the doubling time, and these estimates are shown in Figure 2 against year of publication. The doubling time for citations of Cox (1972) in WoS has considerably increased over the years, meaning that it is now taking considerably longer for the number of citations to double. The doubling time for mentions of the model outlined in Cox (1972) in PM has also increased, but much more slowly, especially after around 2000. This means that the number of PM mentions has been doubling at an almost constant rate since 2000, and this reflects changes in the way that the literature has referred to the model outlined in Cox (1972). Cox's proportional hazards regression has become such a standard method that it is used more and more without a formal reference or citation!

### Conclusion

Providing a full picture of the extent and influence of David Cox's publications is a huge and ongoing task. Here, we have provided some quantification of the influence of his work through the lens of citations and mentions in the scientific literature. David's contributions to the theory and application of statistics are firmly established as essential tools in many areas of knowledge discovery and will no doubt continue to play a vitally important role in years to come. ■



**Figure 1:** Cumulative number of citations against year for David Cox's eight most cited papers. The year of publication is also shown.



**Figure 2:** The estimated doubling time against year for the numbers of citations (WoS, continuous) and mentions (PM, dashed) for Cox (1972). Year refers to the year of publication of the citing paper in WoS and to the year that the mentioning paper was added to the PM database. A little smoothing has been applied.

**Sir David Cox's publishing career was remarkable for the extent and influence of its contributions, as well as for its duration; his works have been published continually from 1947 to 2021**

#### Disclosure statement

The authors declare no competing interests.