Contents lists available at ScienceDirect

European Journal of Internal Medicine

journal homepage: www.elsevier.com/locate/ejim

Editorial

Medicine before and after David Cox

ARTICLE INFO

Keywords
David cox
Proportional hazards model
Binary logistic regression
survival analysis
History
David Cox

ABSTRACT

Herein we recount the legacy of Sir David Roxbee Cox (15 July 1924 – 18 January 2022) from the perspective of practicing clinicians. His-pioneering work in developing the logistic and Cox proportional hazard regression models revolutionized the analysis and interpretation of categorical and time-to-event survival outcomes in modern medicine. This legacy is an inspiration for all those who follow on Sir David Cox's path.

In this editorial we would like to remember the memory of the recently deceased Sir David Roxbee Cox by reviewing the impact his contributions have had on medicine. Cox's legacy in medicine is easy to trace. It is almost impossible to find an issue of EJIM (or any journal) without articles built on the methods developed by Cox. The Cox proportional hazards model alone has currently received more than 30,000 Pubmed citations and remains a cornerstone of modern health research [1,2]. Medical research methodology can indeed be divided in two periods: before Cox and after Cox.

The logistic regression approach commonly used today in statistics and machine learning to model binary outcomes was invented by Joseph Berkson in 1944 [3] and refined by Cox in 1958 [4]. Binary outcomes are typically categorical variables with two mutually exclusive levels such as yes vs. no, or hospitalized vs. not hospitalized. Before logistic regression, these outcomes were analysed by $\chi2$-tests built on contingency tables, or by various other laborious approaches that could not incorporate in their models multiple predictor variables such as patient age or comorbidities. On the other hand, logistic regression can predict the probability of categorical outcomes as a function of several predictor variables. To name a few examples, the Fine pneumonia severity index, Clinical Index of Stable Febrile Neutropenia (CISNE), and Simplified Pulmonary Embolism Severity Index (sPESI) prognostic scales used for predicting the severity of community-acquired pneumonia, complications of febrile neutropenia in patients with cancer, and mortality in patients with pulmonary thromboembolism, respectively, were all developed using logistic regression [5–7]. Cox also developed multinomial logistic regression in 1966, which allowed outcomes with more than two categories thus greatly increasing the scope and popularity of logistic regression models [8].

Linear regression approaches that used a straight line to model the relationships of predictor variables with an outcome were already established at the end of the 18th century by Adrien-Marie Legendre and Carl Friedrich Gauss [9]. However, linear regression does not perform well with categorical outcomes. For example, if the goal is to assess the impact of a continuous numerical predictor (age) on a binary dichotomous outcome such as the presence or absence of complications (represented mathematically by 0 and 1), the predictions would not be bounded within the interval between 0 and 1. In his highly influential 1958 paper [4], Cox suggested the use of the non-linear logistic curve instead of the straight line used in linear regression. The logistic respects the boundaries of the categorical outcome and can be interpreted by a parameter called the odds ratio (OR), which is the ratio of the odds of the outcome (e.g., presence of complications) given one value of a predictor variable (e.g., age = 40 years old) divided by the odds of the outcome given another value of the predictor variable (e.g., age = 41 years old). The odds of the outcome are the probability (P) that the outcome will occur divided by the probability (1-P) that it will not occur [10]. This intuitive interpretation of ORs allowed medical researchers to meaningfully estimate the strength of the association between different predictor variables and categorical outcomes [10,11] Fig. 1. shows a concrete example based on the esophagogastric cancer registry AGAMENON_SEOM in which the binary logistic model allows the proper estimation of the probability of a binary variable whereas linear regression violates the axiom that probabilities cannot have values greater than 1.0 [11,12]. The advent of modern computers facilitated the use of logistic regression models for predicting two or more categorical outcomes. In 2021 alone, >49,000 PubMed articles used a logistic regression model, and this approach continues to often perform favorably compared to more modern machine learning techniques [13,14].

But it was the invention by Sir David Cox of another approach, known as the Cox proportional hazards regression model, that revolutionized the design and interpretation of clinical trials in modern medicine [1]. The growing popularity of prospective randomized clinical trials (RCTs) in the 1970s increased the use of datasets with censored survival data whereby the event of interest, e.g., patient hospitalization, had not yet occurred for a particular patient during their follow-up time. The Kaplan-Meier method, described in 1958, allowed the estimation of

https://doi.org/10.1016/j.ejim.2022.02.022
Received 11 February 2022; Received in revised form 22 February 2022; Accepted 25 February 2022
Available online 28 February 2022

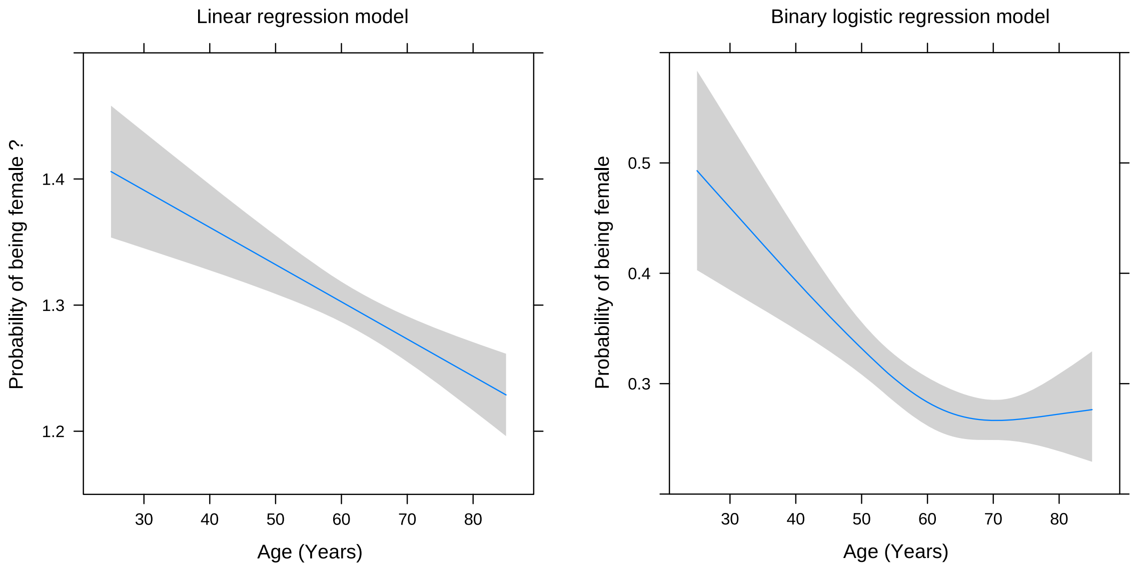

Fig. 1. Comparison of the binary logistic model and linear regression in estimating the probability of a dichotomous variable. The linear regression model wields invalid probabilities greater 1.0.

such time-to-event outcomes in the presence of censored data [15]. The logrank test developed by Nathan Mantel in 1966 further allowed the comparison of such time-to-event outcomes when produced by two groups such as the treatment group versus a control group [16]. The logrank test could not be used to model the effect of multiple predictor variables on the time-to-event outcomes. However, in the same way that logistic regression solved the problem of incorporating multiple predictor variables that χ2-tests could not do, the Cox proportional hazards regression approach allowed the use of multiple predictor variables when comparing time-to-event outcomes that include censored data [17]. Similarly to how the logistic regression yielded the OR, the Cox proportional hazards regression allowed the estimation of a parameter called the hazard ratio (HR), i.e., the ratio of the hazard rates between different values of the predictor variables. A hazard rate represents at a particular time the instantaneous rate of a patient experiencing an outcome, such as death in an overall survival analysis, given that the patient has not yet experienced the event. The HR obtained from a Cox proportional hazards regression model is the ratio of the hazard rate for the outcome given one value of a predictor variable (e.g., age = 40 years old) divided by the hazard rate of the outcome given another value of the predictor variable (e.g., age = 41 years old). The time-to-event outcomes are most commonly assumed to change exponentially over time, and this corresponds to a hazard rate that stays constant at all times for each value of a predictor variable.

The Cox proportional hazards regression model subsequently became the most popular approach for the analysis of survival data in medicine [1]. This was facilitated by Frank Harrell's introduction in 1979 of PHGLM [18], which was the first Cox proportional hazards procedure for the statistical package SAS and a precursor to the contemporary PHREG procedure. Harrell also wrote LOGIST, the first logistic regression procedure for SAS [19], and was thus instrumental in popularizing both of Sir David Cox's major methodological contributions to medical research. For these, and many other scientific contributions, Cox was knighted by Queen Elizabeth II in 1985 and was the first recipient in 2017 of the International Prize in Statistics, considered to be the equivalent of the Nobel Prize for statisticians.

The Cox regression approach was further developed by Cox himself and many others who followed to accommodate more complex scenarios such as cases when the predictor variables change over time [20], or when the proportional hazards assumption does not hold [21]. In medical applications, predictor variables that change over time can introduce a bias called guarantee-time bias, also known as immortal time bias, which may be the single most common cause of invalid survival analyses in the medical literature [22,23]. Cox regression models that account for such time-dependent effects are powerful tools that can be used to avoid the distorting effects of guarantee-time bias [22]. Another key issue in medical applications is scenarios where the proportional hazards assumption does not hold, as is commonly observed in survival analyses of immune checkpoint therapies in oncology [24,25]. A number of solutions have been proposed for these situations as the literature on the topic continues to evolve [21,26–28].

On a philosophical level, Cox took a pragmatic and non-dogmatic view in the 20th century debates between the frequentist and Bayesian schools of statistics, noting that each approach has its own advantages and can be useful in solving practical problems: "*At a more theoretical level it is always instructive to compare solutions of related problems obtained by different approaches or under somewhat different assumptions. From this point of view, Bayesian theory is both interesting and valuable*" [29]. The Bayesian approach that Cox is referring to here allow the estimation of the probabilities, known as posterior probabilities, of individual events such as the posterior probability that a patient has breast cancer if she has a positive mammogram. Bayesian posterior probabilities are naturally applicable to clinical decision-making because clinicians and patients can directly use them when considering risk-benefit trade-offs [11]. Estimation of posterior probabilities always requires us to first specify what the prior probability is, which is a complex task [30]. Since prior probabilities are assumptions, they can lead to invalid or misleading conclusions when they are misspecified [31]. Conversely, the frequentist approach at its most pure form denies any meaning to probabilities of individual events and uses only probabilities that are relative frequencies of events, such as the proportion of women sampled from the same population as our patient who were diagnosed with breast cancer after having a positive mammogram [32]. Bayesian and frequentist approaches are valuable complementary methods of inference that can help us analyze concrete clinical problems [33]. In his practice, Cox mainly used frequentist approaches but was open to Bayesian decision-making if needed. This ecumenical approach allowed him to note key limitations of the frequentist Neyman-Pearson decision-making system: "*… I felt that various aspects of the Neyman-Pearson theory (choose alpha, choose a critical region, reject or accept the null hypothesis, give a rigid procedure), that this isn't the way to do science*" [34]. His-inclusive and pragmatic philosophy was summarized in his 1977 lecture to the Statistical Society of Australia, which called for an eclectic view enriched by knowledge gained from each of what he categorized as the three broad approaches to statistical inference: sampling theory (frequentism), pure likelihood, and Bayesian [35].

In terms of his human values, Sir David Cox was known for his intellectual humility, as often recounted by all of our colleagues who have had the honor of meeting him. He often recalled his errors of judgement, or the times he had given up on solving a problem [34]. In his old age, he wondered whether he had tried hard enough according to his abilities. David Cox had tackled a wide range of problems, from industrial applications to theoretical work, in disciplines ranging from psychology to social science, astronomy and physics. His-work helped catalyze all these disciplines. He wrote >300 articles and >20 books and single-handedly changed the course of medicine. For practicing clinicians, the legacy of Sir David Cox is the history of our profession. Knowing it allows us to understand our present and steadfastly build our future with humility and scientific rigor.

References

[1] Cox DR. Regression models and life-tables. J R Stat Soc Ser B 1972;34:187–202.

[2] van Houwelingen HC, Cook RJ, Joly P, Martinussen T, Taylor JMG, Abrahamowicz M, et al. Analysis of time-to-event for observational studies: guidance to the use of intensity models. Stat Med 2020;40:185–211.

[3] Berkson J. Application of the logistic function to bio-assay. J Am Stat Assoc 1944; 39:357–65.

[4] Cox DR. The regression analysis of binary sequences. J R Stat Soc Ser B 1958;20: 215–32.

[5] Jiménez D, Aujesky D, Moores L, Gómez V, Lobo JL, Uresandi F, et al. Simplification of the pulmonary embolism severity index for prognostication in patients with acute symptomatic pulmonary embolism. Arch Intern Med 2010;170:1383–9. https://doi.org/10.1001/archinternmed.2010.199.

[6] Carmona-Bayonas A, Jiménez-Fonseca P, Echaburu JV, Antonio M, Font C, Biosca M, et al. Prediction of serious complications in patients with seemingly stable febrile neutropenia: validation of the clinical index of stable febrile neutropenia in a prospective cohort of patients from the FINITE study. J Clin Oncol 2015;33:465–71. https://doi.org/10.1200/JCO.2014.57.2347.

[7] Aujesky D, Fine MJ. The pneumonia severity index: a decade after the initial derivation and validation. Clin Infect Dis 2008;47:S133–9.

[8] Cox DR. Some procedures connected with the logistic qualitative response curve. Research papers in probability and statistics. London: Wiley; 1966.

[9] Stigler SM. Gauss and the invention of least squares. Ann Stat 1981;9:465–74.

[10] Hosmer Jr DW, Lemeshow S, Sturdivant RX. Applied logistic regression. vol. 398. London: John Wiley & Sons; 2013.

[11] Msaouel P, Lee J, Thall PF. Making patient-specific treatment decisions using prognostic variables and utilities of clinical outcomes. Cancers (Basel) 2021;13: 2741.

[12] Cotes Sanchís A, Gallego J, Hernandez R, Arrazubi V, Custodio A, Cano JM, et al. Second-line treatment in advanced gastric cancer: data from the Spanish AGAMENON registry. PLoS ONE 2020;15:e0235848.

[13] Ayer T, Chhatwal J, Alagoz O, Kahn Jr CE, Woods RW, Burnside ES. Comparison of logistic regression and artificial neural network models in breast cancer risk estimation. Radiographics 2010;30:13–22.

[14] Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol 2019;110:12–22.

[15] Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc 1958;53:457–81.

[16] Mantel N. Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemother Rep 1966;50:163–70.

[17] Turnbull BW, Brown Jr BW, Hu M. Survivorship analysis of heart transplant data. J Am Stat Assoc 1974;69:74–80.

[18] Harrell FE. The PHGLM procedure. Suppl Libr User's Guid 1983:267–94.

[19] SAS/STAT 13.1 User's Guide The LOGISTIC Procedure n.d. https://support.sas.com/documentation/onlinedoc/stat/131/logistic.pdf (accessed February 6, 2022).

[20] Zhang Z, Reinikainen J, Adeleke KA, Pieterse ME, Groothuis-Oudshoorn CGM. Time-varying covariates and coefficients in Cox regression models. Ann Transl Med 2018;6:121.

[21] Austin PC, Fang J, Lee DS. Using fractional polynomials and restricted cubic splines to model non-proportional hazards or time-varying covariate effects in the Cox regression model. Stat Med 2021;41:612–24.

[22] Giobbie-Hurder A, Gelber RD, Regan MM. Challenges of guarantee-time bias. J Clin Oncol 2013;31:2963.

[23] Kragh Andersen P, Pohar Perme M, van Houwelingen HC, Cook RJ, Joly P, Martinussen T, et al. Analysis of time-to-event for observational studies: guidance to the use of intensity models. Stat Med 2021;40:185–211.

[24] Castañon E, Sanchez-Arraez A, Alvarez-Manceñido F, Jimenez-Fonseca P, Carmona-Bayonas A. Critical reappraisal of phase III trials with immune checkpoint inhibitors in non-proportional hazards settings. Eur J Cancer 2020;136:159–68.

[25] Castañon E, Sanchez-Arraez Á, Jimenez-Fonseca P, Alvarez-Manceñido F, Martínez-Martínez I, Gongora LM, et al. Bayesian interpretation of immunotherapy trials with dynamic treatment effects. Eur J Cancer 2022;161:79–89.

[26] Shepherd BE. The cost of checking proportional hazards. Stat Med 2008;27: 1248–60.

[27] Gray RJ. Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. J Am Stat Assoc 1992;87:942–51.

[28] Brilleman SL, Elci EM, Novik JB, Wolfe R. Bayesian survival analysis using the rstanarm R package. ArXiv Prepr ArXiv200209633 2020.

[29] Cox DR, Hinkley DV. Theoretical statistics. New York: CRC Press; 1979.

[30] Efron B. Modern science and the bayesian-frequentist controversy. Division of Biostatistics, Stanford University; 2005.

[31] Quintana M, Viele K, Lewis RJ. Bayesian analysis: using prior information to interpret the results of clinical trials. JAMA 2017;318:1605–6.

[32] Greenland S. Probability logic and probabilistic induction. Epidemiology 1998: 322–32.

[33] Bendtsen M. A gentle introduction to the comparison between null hypothesis testing and Bayesian analysis: reanalysis of two randomized controlled trials. J Med Internet Res 2018;20:e10873.

[34] Reid N. A conversation with sir david cox. Stat Sci 1994;9:439–55.

[35] Cox DR. Statistical Society of Australia. The Knibbs lecture for 1977. Foundations of statistical inference: the case for eclecticism. Aust J Stat 1978;20:43–59.

Pavlos Msaouel[a,*], Paula Jimenez-Fonseca[b], Bora Lim[c], Alberto Carmona-Bayonas[d], Giancarlo Agnelli[e]

[a] *Department of Genitourinary Medical Oncology, The University of Texas MD Anderson Cancer Center, 1515 Holcombe Boulevard, Houston, TX 77030, United States of America*

[b] *Medical Oncology Department. Hospital Universitario Central de Asturias. Avenida de Roma s/n, Oviedo Asturias. Spain.*

[c] *Breast Oncology, Dan L. Duncan Cancer Center, Baylor College of Medicine, Houston, TX, United States of America*

[d] *Hematology and Medical Oncology Department, Hospital Universitario Morales Meseguer. UMU. IMIB. Murcia. Spain*

[e] *Internal Vascular and Emergency Medicine-Stroke Unit, University of Perugia, Perugia, Italy*

[*] Corresponding author.

E-mail addresses: PMsaouel@mdanderson.org (P. Msaouel), Bora.Lim@bcm.edu (B. Lim).