EUROPEAN JOURNAL OF CANCER 44 (2008) 448–453

available at www.sciencedirect.com

ScienceDirect

journal homepage: www.ejconline.com

Do cancer predictions work?

Tadeusz Dyba, Timo Hakulinen*

Finnish Cancer Registry, Liisankatu 21 B, FI-00170 Helsinki, Finland

ARTICLE INFO

Article history:
Received 22 August 2007
Received in revised form
29 October 2007
Accepted 9 November 2007
Available online 20 December 2007

Keywords:
Cancer
Incidence
Projections
Predictions
Confidence interval
Poisson distribution
Ex post prediction

ABSTRACT

Two different types of simple extrapolation models were investigated as tools for cancer incidence prediction. The number of incident cancer cases by sex and site in Finland was predicted using a prediction interval for each year from 1967 to 2003 on the basis of historical cancer incidence data obtained 5 to 15 years earlier. Cancer sites where major human-made changes in aetiology and diagnostics had occurred were analysed separately. Assuming that such changes had not occurred, the 95% prediction intervals based on normal errors of the age-standardised rate and on Poisson models included the observed number in 65–100% of the years. The Poisson models produced, on average, shorter intervals and were more capable of indicating a site where the model assumptions did not hold true. Simple extrapolation models may be used with some caution on coverage when there are no known factors that might make abrupt changes in the temporal development of cancer incidence. On the other hand, they may be used for detecting the effects of such factors.

1. Introduction

Predictions of occurrence of a disease are mainly made for administrative purposes.[1] These predictions should be as accurate as possible. Using these predictions the authorities could make plans for proper allocation of resources for the control of the disease, its prevention, diagnostics and treatment and for the rehabilitation of the patients. Predictions could also have other purposes, e.g. to show what would be the outcome if a particular programme targeted against the disease were launched or not.

The simplest way to make predictions is to assess the prevailing trends in the occurrence and to extrapolate them into the future. For example, with most cancers, the causes and their distribution in the population are not well-known. With the long latency times between the causes and the effect it is reasonable to assume that the causal effects and their changes would be reflected in the past temporal developments of characteristics of the disease incidence.[1,2] On the other hand, changes in the diagnostic facilities and definitions may create artefactual developments that are not likely to be extrapolatable into the future.[3] Predictions can then also be used to detect such developments.

It is important that simple models are used for extrapolation.[4] Statistical models should be parsimonious, and complicated models are not likely to hold true in the future. Simple models are also easier to interpret and, if valid, guarantee a higher precision for predictions.

It would be important to express the precision of any given prediction. There are three sources of error that may exert an effect on predictions based on extrapolations. First, there is randomness in the historical data on which a model should be fitted. Second, there is inherent randomness in the disease counts to be observed in the future and third, there may be an error due to mis-specification of the model applied. The first two sources of error may be controlled by appropriate prediction intervals,[2] whereas the third source has led to attempts to use Bayesian modelling with credible intervals.[5,6]

* *Corresponding author:* Tel.: +358 9 135 331; fax: +358 9 135 5378.
E-mail address: tadek.dyba@cancer.fi (T. Dyba).
0959-8049/$ - see front matter
doi:10.1016/j.ejca.2007.11.014

Experience with the model may be used as a practical guidance about the use of a particular model. The present paper reports on the success of two alternative ways of making simple linear extrapolations to predict cancer incidence in Finland during the past decades, had both of the current methodologies been available in the past.

2. Materials and methods

2.1. Material

The Finnish Cancer Registry has collected data of high quality on all cancer cases diagnosed in Finland since 1953.[7] The most common sites were included for both sexes, with the exception of five sites, based on a priori knowledge. Cancers of the breast and cervix uteri were excluded, as for them, countrywide mass-screening programmes are ongoing, the features of which may be taken into account when making predictions.[8–10] Changed diagnostic practices have enormously affected the incidence of prostatic cancer,[11] which was excluded. It was also known a priori that dramatic changes had occurred in the smoking habits of males, which could also be taken into account when making predictions.[12] As the changes in women's smoking habits had been slow and smooth,[13] female lung cancer was included. There had also been changes in the definition of urinary bladder cancer and consequently this cancer was also excluded.[14] Finally, the material included 15 sites for males and 18 sites for females. For comparison, analyses were also made for the excluded sites.

As the models were simple, only 10 years of historical data were included in the prediction base, i.e. the data on which the statistical models used were fitted. The horizon of prediction was 5 years in the future for each combination of sex and site. With these choices, the various prediction bases covered the years 1953–1988 and the predictions were made for the 37 single years in 1967–2003.

2.2. Methods

Two different approaches were used, the simple extrapolation of the age-standardised incidence rates[15] and the Poisson regression models.[16,17] The observed population age structure in each year was used as a reference in the standardisation, to make a simple transformation from the age-standardised rate to the number of cases. Within each approach, an automatic model selection algorithm was used. If the slope of overall incidence was positive, the incidence rates or counts were modelled as such, to avoid exponential growths in rates. If it was negative, a logarithmic transformation was employed, to avoid negative predictions. For zero or close to zero slope both alternatives coincided or nearly coincided, and thus the model choice then was not a real issue.

For the Poisson models, within each choice there were two alternatives. For positive slope, they were[2,17]

$$Ec_{it} = n_{it}(\alpha_i + \beta_i t)$$

and

$$Ec_{it} = n_{it}\alpha_i(1 + \beta t),$$

where c_{it} is the number of cases in age group i in year t, n_{it} is the number of person-years in the same stratum and α_i, β_i and β are the model parameters. E is the symbol of expectation according to the model. The former model postulates a simple linear trend for the incidence c_{it}/n_{it} according to time t. The latter model assumes a simplification that the slope $(\alpha_i\beta)$ is proportional to the intercept (α_i) in the linear trend.

For negative slope, the two choices were[2,17]

$$Ec_{it} = n_{it}\exp(\alpha_i + \beta_i t)$$

and

$$Ec_{it} = n_{it}\exp(\alpha_i + \beta t),$$

with similar interpretations for the logarithm of the incidence to those of the two models concerning positive slope.

The model choice between the two alternatives was based on the log-likelihood statistics. A possibility for over-dispersion was also allowed.

Prediction intervals (95% level) were calculated for both approaches.[15,16]

Two different indices of success were employed. First, the proportion of years with the observed number of new cases or the age-standardised rate falling within the prediction interval was calculated for each combination of sex and site. Second, average coefficients of variation (standard deviation/mean) were assessed for the predictions to compare the lengths of the prediction intervals, by sex and site.

3. Results

For a number of sites, the 95% prediction intervals included the observed number of cases in 95% or more of the years, but for a few sites the coverage was as low as 65% (Table 1). The median coverage proportions were 86% for both the standardised rate extrapolation and the Poisson models.

In female lung cancer, the coverage proportion was the same for the age-standardised rate extrapolation and the Poisson models (Fig. 1). The prediction intervals using the Poisson models were, on an average, shorter and their lengths varied less between the different years. The female lung cancer was an extreme example of a general tendency, seen both in males and females, for the length of the prediction interval (divided by the prediction itself) to be shorter for Poisson models than for extrapolations based on the age-standardised rates (Fig. 2).

Male lung cancer was excluded from the analyses because of drastic changes that had occurred in the smoking habits of males in Finland. When the predictions were made for that cancer, the coverage proportion by extrapolation of the age-standardised rates was 62%, whereas for the Poisson models the coverage was only half of that, 31% (Fig. 3). For lung cancer in males, dramatic changes in the smoking habits caused drastic non-linear developments by time. Thus, the assumptions for using simple linear extrapolation models did not hold in practice, and this was reflected by far lower coverage proportions than 95%. The standardised rate extrapolation method produced, on an average, longer prediction intervals than the Poisson models and thus their coverage of the observed numbers was higher.

Table 1 – Empirical coverage (in %) of the ex post 95% prediction intervals for the annual number of new cases of cancer diagnosed in Finland, in 1967–2003, horizon of prediction = 5 years, by sex, site and method

Site	Females		Males	
	Poisson models	Standardised rate extrap. models	Poisson models	Standardised rate extrap. models
1. Lip	95	95	95	95
2. Oesophagus	95	92	95	92
3. Stomach	84	84	100	100
4. Colon	97	97	73	86
5. Rectum	89	95	97	95
6. Liver	78	97	65	81
7. Gallbladder	76	84	65	81
8. Pancreas	89	89	97	97
9. Lung	84	84	–	–
10. Corpus uteri	86	91	–	–
11. Ovary	86	78	–	–
12. Kidney	84	78	78	81
13. Skin melanoma	89	86	86	92
14. Skin non-melanoma	68	72	81	81
15. Nervous system	76	76	78	95
16. Thyroid	68	68	86	92
17. Non-Hodgkin	85	81	86	92
18. Leukaemia	81	81	92	84

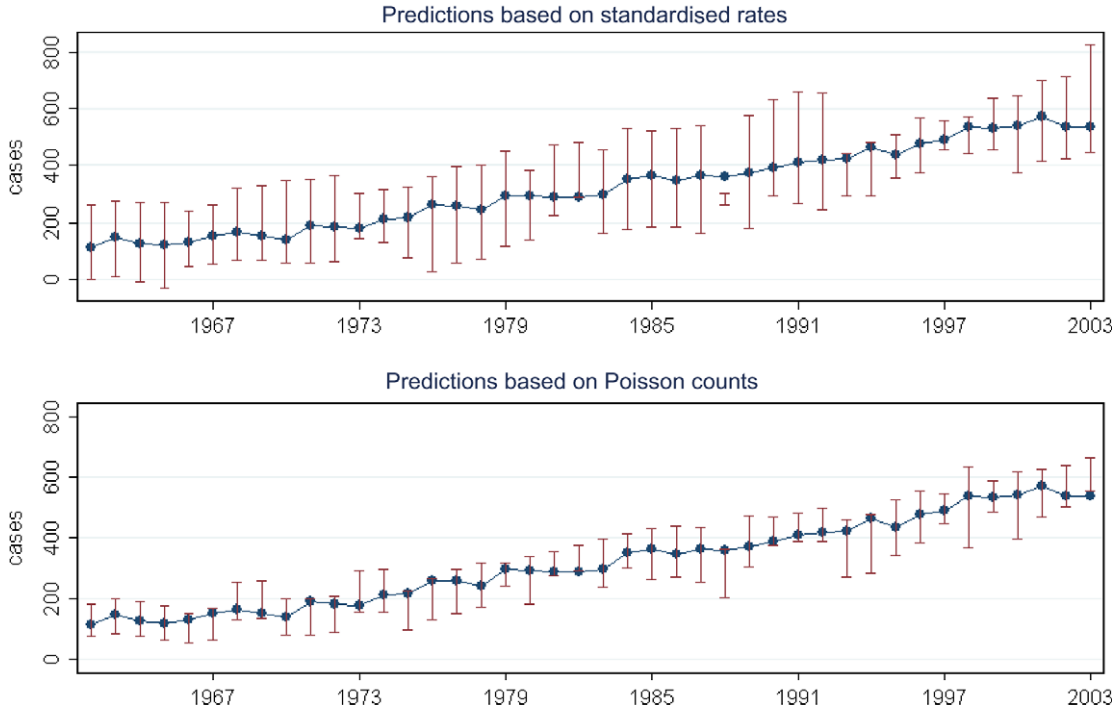

Fig. 1 – Empirical coverage of the ex post 95% prediction intervals, by method, for the annual number of new lung-cancer cases diagnosed in females in Finland in 1967–2003, with a 10-year base and a 5-year horizon of prediction.

4. Discussion

For most cancer sites, it is reasonable to attempt short-term predictions by trend extrapolation, as the causes of cancer and their distribution in the population are not sufficiently known. The process can be automatised for these sites and the coverage proportions of the prediction intervals, even though they are often smaller than the nominal levels, give

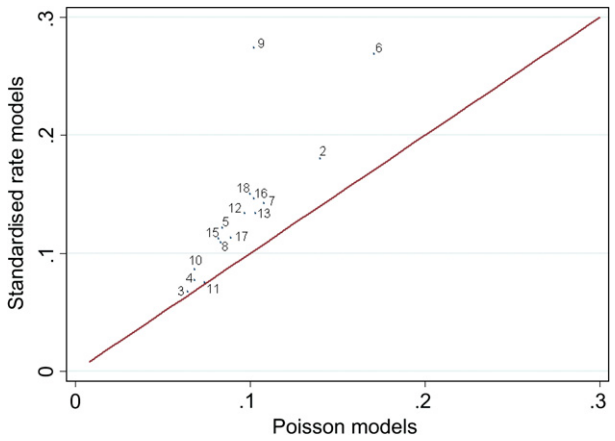

Fig. 2 – Average coefficients of variation for predictions by cancer site and method, with a 10-year base and a 5-year horizon of prediction, females. The different number codes indicate the sites, cf. Table 1.

a better indication of uncertainty than the sole point predictions. If the prediction interval is long or wide, the prediction is imprecise.

An important special use of the prediction is the assessment of the number of new cancer cases in the current year.[18,19] A typical cancer registry publishes the statistics after 2–4 years of the current year. When a 3-year horizon of prediction was used the coverage proportions were, as a rule, close to the nominal values.

It is important to stress that the simple extrapolation models should only be used when the model assumptions are valid. The smooth developments in disease incidence are typically disturbed by activities targeted towards preventing the cancer (e.g. anti-smoking campaigns or early detection of precancerous lesions) or cancer death (mammographic screening of breast cancer or PSA diagnostics of prostatic cancer). Application of simple extrapolation models is supposed to fail in these cases, and other models should be used where these human actions are explicitly taken into account. Thus, the applicability of the prediction intervals must only be evaluated for sites where the simple extrapolations are expected to work.[20]

In this study, two types of models were applied. On an average, the extrapolation of the age-standardised incidence based on normality of random error resulted in longer and more variable prediction intervals than the use of Poisson models. As in lung cancer for males they were also, consequently, often less able to reveal that the model assumptions did not hold.

Predictions may actually be used to show the effect of a human-made programme against cancer. A simple extrapolation prediction will indicate the range of outcomes if a recent trend development held true in the future. A success of the programme can be demonstrated by the observed number of cases being below the prediction interval.

The ex post predictions were assessed with a 100% success of the population forecasts, i.e. the predictions were made only after the exact population count was known. For the age-standardised rates the success or failure of the prediction is not an issue as these rates are unchanged by population numbers. For numbers of new cases, however, it is also very crucial that the population forecasts work. Experience with the standardised rate extrapolation models showed that many of the predictions for the numbers of new cases could

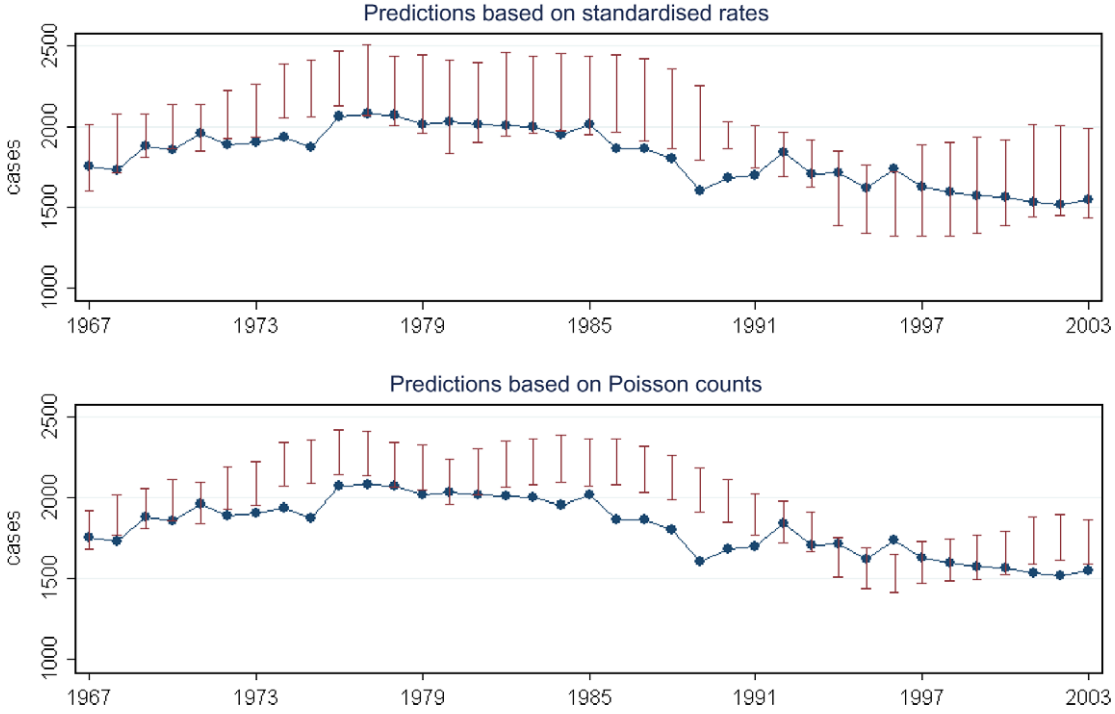

Fig. 3 – Empirical coverage of the ex post 95% prediction intervals, by method, for the annual number of new lung-cancer cases diagnosed in males in Finland in 1967–2003, with a 10-year base and a 5-year horizon of prediction.

fail in spite of the success of the corresponding prediction for the age-standardised rate.[15]

The fact that in this study the coverage proportions quite often were not 95% means that, to some extent, the sources of error other than the simple Poisson error have also had an effect on the observed rates. Therefore, it is not sufficient to only show the theoretical performance of the method with simulated data[17] but a validation with real data is also important.

Due to randomness, the coverage proportion will never be exactly 95%, even if only Poisson variation affected the incidence rates. When there were changes in the trends, e.g. in melanoma of the skin due to better awareness of the risks related to sunbathing,[21] the prediction intervals could be expected to cover less than 95% of the observed rates. The coverage proportions were, however, close to 90%, indicating that the changes were almost smooth enough, on an average, to be picked up with the prediction method based on a short, 10-year, prediction base. The changes in non-melanoma skin cancer were much more abrupt and the coverage proportions were much lower.

It is difficult to take proper account of model mis-specification. Bayesian models with very wide (almost incredible) credible intervals have been attempted.[5,6] Another possibility is to try to explicitly incorporate into the model the factors (e.g. smoking, mass-screening) that may help the model from being mis-specified.[10,22,23]

It is useful to try to predict the future incidence rates with a number of differently specified models.[24] Particularly helpful is the consideration of predictions based on different scenarios of future developments.[10,12,22,23]. This will not only be so for the possible model mis-specification but also will help in understanding different consequences of human actions. This approach may also give a judicious choice of prediction approach when the rates are changing rapidly. Another possibility are the age-period-cohort models.[24] Prediction intervals can, for all of these predictions, be constructed using the approach outlined by Hakulinen and Dyba.[16] This method is general and by no means restricted to age-period models only.

Bayesian[5,6] and Frequentist age-period-cohort models[24] require a longer historical time-series as a basis of prediction. Cohort phenomena are not unknown in the temporal analysis of cancer incidence rates[25] but they require a longer time span to be detected. On the other hand, for predictions, it does not often matter even if there are rate developments by birth cohort, as long as they are linear on absolute or logarithmic scales. The non-identifiability property of the models means that linear developments by birth cohort may be reformulated as linear developments by calendar time,[2] and thus the prediction can be reformulated as a simple trend extrapolation by calendar time. Thus, the use of very long historical time series as prediction basis may not always be warranted. Doll and Peto[26] have thoroughly discussed the issue of the sources of bias in estimating trends in cancer incidence. There is another problem with longer time-series of incidence data, relating to the stability of diagnostic practices and definitions over time. With the models, it may be risky to extrapolate developments that essentially already took place a long time ago.

The simple extrapolation models can also be easily generalised to incorporate other factors, e.g. region or other definition of a population group. Trend analysis within a country may well show that there is a general tendency of incidence development but the level of incidence varies by region. This kind of pattern can be handled by simply adding a region term in the models.

Predicting a future incidence of cancer is a process which, for the main large sites of cancer, lung, prostate and breast, may not be properly handled by simply extrapolating the recent cancer incidence rates. But for the majority of the sites, extrapolation may well be the method of choice.

Conflict of interest statement

None declared.

Acknowledgement

This study was supported by the Cancer Society of Finland and the Academy of Finland, MaDaMe project.

REFERENCES

1. Hakama M, Hakulinen T. Trends and projections in cancer risk in Finland. In: Balducci L, Lyman GH, Ershler WW, editors. *Geriatric oncology*. Philadelphia: Lippincott; 1992. p. 370–4.
2. Dyba T, Hakulinen T, Päivärinta L. A simple non-linear model in incidence prediction. *Stat Med* 1997;**16**:2297–309, 2000;**19**:1251.
3. Saxén EA. Trends: facts or fallacy. In: Magnus K, editor. *Trends in cancer incidence. Causes and practical implications*. New York: Hemisphere; 1982. p. 5–16.
4. Chatfield C. *Time-series forecasting*. Boca Raton: Chapman & Hall; 2001.
5. Bashir SA, Estève J. Projecting cancer incidence and mortality using Bayesian age-period-cohort models. *J Epidem Biostat* 2001;**6**:287–96.
6. Bray I. Application of Markov Chain Monte Carlo methods to projecting cancer incidence and mortality. *Appl Stat* 2002;**51**:151–64.
7. Teppo L, Pukkala E, Lehtonen M. Data quality and quality control of a population-based cancer registry. Experience in Finland. *Acta Oncol* 1994;**33**:365–9.
8. Hakama M, Räsänen-Virtanen U. Effect of a mass screening program on the risk of cervical cancer. *Am J Epidemiol* 1976;**103**:512–7.
9. Møller B, Weedon-Fekjær H, Hakulinen T, et al. The influence of mammographic screening on national trends in breast cancer incidence. *Eur J Cancer Prev* 2005;**14**:117–28.
10. Seppänen J, Heinävaara S, Hakulinen T. Influence of alternative mammographic screening scenarios on breast cancer incidence predictions (Finland). *Cancer Causes Control* 2006;**17**:1135–44.
11. Møller B, Fekjær H, Hakulinen T, et al. Prediction of cancer incidence in the Nordic countries up to the year 2020. *Eur J Cancer Prev* 2002;**11**(Suppl. 1).
12. Hakulinen T, Pukkala E. Future incidence of lung cancer: forecasts based on hypothetical changes in the smoking habits of males. *Int J Epidemiol* 1981;**10**:233–40.

13. Vartiainen E, Jousilahti P, Alfthan G, Sundvall J, Pietinen P, Puska P. Cardiovascular risk factor changes in Finland, 1972–1997. *Int J Epidemiol* 2000;**29**:49–56.
14. Silverman DT, Devesa SS, Moore LE, Rothman N. Bladder cancer. In: Schottenfeld D, Fraumeni Jr JF, editors. *Cancer epidemiology and prevention*. Oxford: Oxford University Press; 2006. p. 1101–27.
15. Hakulinen T, Teppo L, Saxén E. Do predictions for cancer incidence come true? Experience from Finland. *Cancer* 1986;**57**:2454–8.
16. Hakulinen T, Dyba T. Precision of incidence predictions based on Poisson distributed observations. *Stat Med* 1994;**13**:1513–23.
17. Dyba T, Hakulinen T. Comparison of different approaches to incidence prediction based on simple interpolation techniques. *Stat Med* 2000;**19**:1741–52.
18. Boring CC, Squires TS, Tong T, Montgomery S. Cancer statistics 1994. *CA Cancer J Clin* 1994;**44**:7–26.
19. Finnish Cancer Registry. Institute for Statistical and Epidemiological Cancer Research. Cancer in Finland 2002 and 2003. Helsinki: Cancer Society of Finland Publ No. 66; 2005.
20. Bray F, Møller B. Predicting the future burden of cancer. *Nature Rev Cancer* 2006;**6**:63–74.
21. Stang A, Stabenow R, Eisinger B, Söderman B, Hakulinen T. Site and sex-specific time trend analyses of the skin melanoma incidence in the former German Democratic Republic including 19,531 cases. *Eur J Cancer* 2003;**39**:1610–8.
22. Brown CC, Kessler LG. Projections of lung cancer mortality in the United States 1985–2025. *J Natl Cancer Inst* 1988;**80**:43–51.
23. White E. Projected changes in breast cancer incidence due to the trend toward delayed childbearing. *Am J Public Health* 1987;**77**:495–7.
24. Møller B, Fekjær H, Hakulinen T, et al. Prediction of cancer incidence in the Nordic countries; empirical comparison of different approaches. *Stat Med* 2003;**22**:2751–66.
25. Coleman MP, Estève J, Damiecki P, Arsan A, Renard H. Trends in cancer incidence and mortality. IARC Scientific Publ No. 121. Lyon; 1993.
26. Doll R, Peto R. The causes of cancer: quantitative estimates of avoidable risks of cancer in the United States today. *J Natl Cancer Inst* 1981;**66**:1191–308.