American Journal of Epidemiology
Published by Oxford University Press on behalf of the Johns Hopkins Bloomberg School of Public Health 2021.

Vol. 191, No. 3
https://doi.org/10.1093/aje/kwab285
Advance Access publication:
December 7, 2021

Invited Commentary

Invited Commentary: Predicting Incidence Rates of Rare Cancers—Adding Epidemiologic and Spatial Contexts

Ian D. Buller and Rena R. Jones*

* Correspondence to Dr. Rena Jones, Occupational and Environmental Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, 9609 Medical Center Drive, Room 6E606, Rockville, MD 20850 (e-mail: rena.jones@nih.gov).

Initially submitted October 8, 2021; accepted for publication November 2, 2021.

There are unique challenges to identifying causes of and developing strategies for prevention of rare cancers, driven by the difficulty in estimating incidence, prevalence, and survival due to small case numbers. Using a Poisson modeling approach, Salmerón et al. (*Am J Epidemiol.* 2022;191(3):487–498) built upon their previous work to estimate incidence rates of rare cancers in Europe using a Bayesian framework, establishing a uniform prior for a measure of variability for country-specific incidence rates. They offer a methodology with potential transferability to other settings with similar cancer surveillance infrastructure. However, the approach does not consider the spatiotemporal correlation of rare cancer case counts and other, potentially more appropriate nonnormal probability distributions. In this commentary, we discuss the implications of future work from cancer epidemiology and spatial epidemiology perspectives. We describe the possibility of developing prediction models tailored to each type of rare cancer; incorporating the spatial heterogeneity in at-risk populations, surveillance coverage, and risk factors in these predictions; and considering a modeling framework with which to address the inherent spatiotemporal components of these data. We note that extension of this methodology to estimate subcountry rates at provincial, state, or smaller geographic levels would be useful but would pose additional statistical challenges.

incidence rate; prediction; rare cancers; spatial epidemiology

Abbreviations: INLA, integrated nested Laplace approximation; RARECARENet, Information Network on Rare Cancers.

Editor's note: *The opinions expressed in this article are those of the authors and do not necessarily reflect the views of the* American Journal of Epidemiology.

In their accompanying article in this issue of the *Journal*, Salmerón et al. (1) evaluated a Poisson modeling approach for estimating incidence rates of rare cancers (defined as an incidence of 6 or fewer cases per 100,000 individuals per year), building on their earlier work (2), which suggested shortcomings of integrated nested Laplace approximation (INLA) for estimating random effects with Poisson-distributed data. The authors used 2 studies to establish a uniform hyperprior distribution for the precision parameter within the Poisson model ($\sigma \sim \text{Uniform}(1, 500)$) as a reliable measure of variability for country-specific incidence rates in a Bayesian modeling framework. Specifically, the 2 studies were designed to 1) evaluate the relative performance of multiple frequentist, empirical Bayes, and Bayesian approaches in predicting rare cancer incidence rates and 2) determine the parameterization for a uniformly distributed hyperprior in the best-fitting model. The authors also proposed a model selection paradigm that compares a suite of model fit indicators. Using data from a European Union coalition of 94 population-based cancer registries called the Information Network on Rare Cancers (RARECARENet), Salmerón et al. predicted incidence rates for 190 rare cancers in 27 European countries (1).

This work serves to extend the existing methodology to estimate the burden of rare cancers, which is challenging due to small populations, exceptionally rare cancers, and variable quality and coverage in disease surveillance systems within and across geographic regions. Rare cancers are more

commonly diagnosed at regional or distant stages, which partially explains their poorer 5-year relative survival in comparison with common cancers (3, 4). Efforts for prevention, early detection, and better understanding of the etiology of these cancers are directly impacted by the ability to accurately estimate incidence, prevalence, and survival. The small numbers of these cancers can result in model estimates that are based on unstable empirical data, rendering them less intuitive for interpretation. The authors raise several focal points for future work that are worth emphasis and elaboration here.

There is likely no one-size-fits-all model specification for the 190 rare cancers within RARECARENet, because the at-risk population and risk factors are spatially heterogeneous and unique for each cancer. Population density varies widely across Europe, from rural areas of Scandinavia (<35 persons/km^2) to bustling metropolises like Paris, France (21,000 persons/km^2) (5). Smoking is a known carcinogen and a risk factor for many cancers (6, 7), and smoking rates also vary widely across Europe (8). Levels of environmental carcinogens such as outdoor air pollutants (9) and radon (10), as well as certain occupational exposures (11), vary geographically across Europe. Finally, over 200 cancer registries within the European Network of Cancer Registries compile cancer incidence counts. While the European Network of Cancer Registries covers most of the European population (12), registration is variable across European countries, and geographic gaps in coverage (and data availability) exist (13). Salmerón et al. (1) did not consider these factors in their model specification, which could improve incidence rate predictions for cancers with small incidence counts. Investigators may consider prediction models tailored to each type of rare cancer as an alternative approach.

The spatial heterogeneity in the at-risk population, potential exposures, and surveillance coverage should be considered when estimating both rare and common cancers. However, major cancer sites with large incidence counts do not tend to suffer from the small numbers characteristic of rare cancers. The incidence rates of common cancers with large counts are often predicted as the expected value given the at-risk population, incidence counts, and age distributions at multiple time points. These efforts use frequentist methods such as age-period-cohort models (14, 15), joinpoint regression (16), or a suite of averaging techniques based on cancer registry coverage within countries or neighboring countries (17). Prediction of incidence rates at major cancer sites benefits from incorporation of potential risk factors for incidence in the predictions (18), data availability permitting.

Although it was not attempted by Salmerón et al. (1), small counts of rare cancers present a unique challenge for predicting cancer incidence rates at subcountry levels (e.g., provinces, states, districts, counties), which may be important for etiological research on hypothesized (localized) risk factors. Here, the challenge of small numbers would be further compounded by small spatial areas that present concerns about data privacy and statistical instability. Small counts in small areas increase the chance of personal identification of cancer cases. In the United States, the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) program (www.seer.cancer.gov), a national source for historical population-based cancer incidence data, protects the privacy of cancer survivors by censoring the release of data for counties with fewer than 16 cases (19). These protections generate gaps in the cancer incidence data and present an opportunity to predict these rates instead. The possibility of subcountry-level predictions will also depend on the rarity of the cancer, the size of the at-risk population, and the temporal data range and granularity of case counts, which may not be available or consistently collected for all subdivisions of a country.

Geography is not only a source of variation in the cancer rates but also a potential source of their correlation. The authors designed their study in the presence of overdispersion, where the data have a larger variance than expected from the probability distribution (20). Fitting a model without considering the presence of overdispersion may lead to improper interpretation of model coefficients because of the underestimation of standard errors of estimated coefficients (20). Overdispersion can be caused by the presence of unconsidered correlation in the data (21), such as spatial autocorrelation, where incidence rates in nearby countries are more similar than incidence rates in countries more geographically distant (22). While Salmerón et al. did not conduct spatial analyses, they recommended a spatial framework for a future study (1). Such a study could consider a separate spatial model for each rare cancer in RARECARENet to further account for overdispersion. Temporal correlation of incidence rates may also be a source of overdispersion, and investigators can also consider spatiotemporal model specifications (18, 23).

It may be possible to extend Salmerón et al.'s analysis to a spatial framework. A popular disease mapping approach is a lognormal Poisson spatial model, such as a Besag-York-Mollié model (24), with a spatially structured residual and an intrinsic conditional autoregressive structure that assumes spatial adjacency of the European countries (25). Other spatial and spatiotemporal model specifications that account for overdispersion include spatial conditional models (26), zero-inflated Poisson models (27), hurdle models (27, 28), and joint models (29). All approaches are implementable in the software packages OpenBUGS (30) and R-INLA (31), used by the authors (26–29, 32, 33). Some technical challenges exist, including a high computational cost for these models in OpenBUGS and known limitations of R-INLA implementation. Salmerón et al. noted the sizable computational burden of a Bayesian approach using Markov chain Monte Carlo sampling as compared with the approximate Bayesian approach with INLA (1). While INLA may perform more quickly than OpenBUGS, complicated model specifications or new prior distributions dependent on the nature of a specific rare cancer are not available within R-INLA (34).

Appropriate choice of the probability distribution is another way to address overdispersion, and model specification will depend on the distribution of case counts. Rare cancer case counts may not follow a Poisson distribution. Salmerón et al. did not test other probability distributions in their frequentist or Bayesian model simulations (1). Their choice of empirical Bayes approach (their model 10), however,

assumes that case counts follow a negative binomial distribution. The negative binomial distribution is commonly used as a model for overdispersed case-count data (35), and a negative binomial regression could be specified in a Bayesian framework to compare with the selected Poisson model. If the true cancer incidence rate is nearly 0, a more appropriate model specification may be a nonnormal distribution such as, for example, a zero-inflated Poisson (36) or hurdle (37, 38) model. In future efforts, investigators could conduct simulation studies and empirical sensitivity analyses of various probability distributions of case counts to predict incidence rates of rare cancers.

One of the goals of the work of Salmerón et al. was to develop a methodology that could be applied across European countries, despite their different distributions of rare cancers. A one-size-fits-all approach has some clear advantages, including the transferability of the methodology to other settings with similar surveillance infrastructure. As an example, the US Surveillance, Epidemiology and End Results program (39) has been predicting incident cancer rates and counts for decades (40), often using a weighted, 2-dimensional, median-based smoothing algorithm called "head-banging" (41) to stabilize rates for areas with low case counts, sparse populations, or both. The methodology proposed by Salmerón et al. could potentially be used to predict rare cancer incidence rates in the United States and globally, especially for programs that typically limit their predictions to cancer sites with large case counts. The American Cancer Society predicts US cancer incidence rates for 47 specific cancer sites, and selected rare cancers are combined, not modeled separately (23, 42). The World Health Organization's GLOBOCAN 2020 database did not predict rare cancers but instead reported on 36 major cancer sites for 185 countries or territories with at least 150,000 persons in 2020 (43). The methods presented by Salmerón et al. could be considered part of an ensemble of approaches for predicting rare cancer incidence rates after the appropriate assumptions and other considerations have been made.

There are unique challenges to identifying causes of and developing strategies for prevention of rare cancers, and ascertaining their burden is an important first step. The work of Salmerón et al. (1) offers a robust methodology for estimating the incidence of rare cancers, with potential transferability to other settings with similar infrastructure for cancer surveillance. In future work, researchers could evaluate models with spatiotemporal components and other nonnormal probability distributions for case counts, both of which are likely relevant to rare outcomes. Extension of this methodology to estimate subcountry-level rates at provincial, state, or smaller geographic levels would support the exploration of specific etiological hypotheses but would present difficult statistical challenges.

ACKNOWLEDGMENTS

Author affiliations: Occupational and Environmental Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, Maryland, United States (Ian D. Buller, Rena R. Jones); and Cancer Prevention Fellowship Program, Division of Cancer Prevention, National Cancer Institute, Rockville, Maryland, United States (Ian D. Buller).

This work was funded by the Intramural Research Program of the National Cancer Institute.

We thank Drs. Barry Graubard and Mary Ward for their thoughtful comments and suggestions.

The opinions expressed by the authors are their own, and this material should not be interpreted as representing the official viewpoint of the US Department of Health and Human Services, the National Institutes of Health, or the National Cancer Institute.

Conflict of interest: none declared.

REFERENCES

1. Salmerón D, Botta L, Martínez JM, et al. Estimating country-specific incidence rates of rare cancers: comparative performance analysis of modelling approaches using European cancer registry data. *Am J Epidemiol.* 2022; 191(3):487–498.
2. Botta L, Capocaccia R, Trama A, et al. Bayesian estimates of the incidence of rare cancers in Europe. *Cancer Epidemiol.* 2018;54:95–100.
3. DeSantis CE, Kramer JL, Jemal A. The burden of rare cancers in the United States. *CA Cancer J Clin.* 2017;67(4):261–272.
4. American Cancer Society. Special section: rare cancers in adults. In: *Cancer Facts & Figures 2017*. Atlanta, GA: American Cancer Society; 2017:30–39. https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2017/cancer-facts-and-figures-2017-special-section-rare-cancers-in-adults.pdf. Accessed October 7, 2021.
5. Eurostat. Population statistics at regional level. (Eurostat population projections 2019). https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Population_statistics_at_regional_level. Published March 2021. Accessed September 29, 2021.
6. Thun MJ, Linet MS, Cerhan JR, et al. *Schottenfeld and Fraumeni Cancer Epidemiology and Prevention*. 4th ed. New York, NY: Oxford University Press; 2018.
7. IARC Working Group on the Evaluation of Carcinogenic Risks to Humans. *Tobacco Smoke and Involuntary Smoking*. (IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, vol. 83). Lyon, France: International Agency for Research on Cancer; 2004.
8. Zatoński W, Przewoźniak K, Sulkowska U, et al. Tobacco smoking in countries of the European Union. *Ann Agric Environ Med.* 2012;19(2):181–192.
9. IARC Working Group on the Evaluation of Carcinogenic Risks to Humans. *Outdoor Air Pollution*. (IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, vol. 109). Lyon, France: International Agency for Research on Cancer; 2016.
10. Berrington De González A, Bouville A, Rajaraman P, et al. Ionizing radiation. In: Thun MJ, Linet MS, Cerhan JR, et al., eds. *Schottenfeld and Fraumeni Cancer Epidemiology and Prevention*. 4th ed. New York, NY: Oxford University Press; 2018:227–248.
11. Kauppinen T. Occupational exposure to carcinogens in the European Union. *Occup Environ Med.* 2000;57(1):10–18.

12. Siesling S, Louwman WJ, Kwast A, et al. Uses of cancer registries for public health and clinical research in Europe: results of the European Network of Cancer Registries survey among 161 population-based cancer registries during 2010–2012. *Eur J Cancer.* 2015;51(9):1039–1049.
13. Forsea A-M. Cancer registries in Europe—going forward is the only option. *Ecancermedicalscience.* 2016;10:641.
14. Institute of Population-Based Cancer Research, Cancer Registry of Norway. Nordpred: prediction of cancer incidence in the Nordic countries up to the year 2020. https://www.kreftregisteret.no/en/Research/Projects/Nordpred/. Accessed October 1, 2021.
15. Dyba T, Hakulinen T. Comparison of different approaches to incidence prediction based on simple interpolation techniques. *Stat Med.* 2000;19(13):1741–1752.
16. Kim HJ, Fay MP, Feuer EJ, et al. Permutation tests for joinpoint regression with applications to cancer rates. *Stat Med.* 2000;19(3):335–351.
17. Ferlay J, Colombet M, Soerjomataram I, et al. Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. *Int J Cancer.* 2019;144(8):1941–1953.
18. Pickle LW, Hao Y, Jemal A, et al. A new method of estimating United States and state-level cancer incidence counts for the current calendar year. *CA Cancer J Clin.* 2007;57(1):30–42.
19. National Center for Chronic Disease Prevention and Health Promotion, Centers for Disease Control and Prevention. *National Program of Cancer Registries Cancer Surveillance System (NPCR-CSS). 2018 Data Release Policy. Diagnosis Years 1995–2017*. Atlanta, GA: Centers for Disease Control and Prevention; 2018. (OMB report 0920-0469). (ICR report 201908-0920-003). https://omb.report/icr/201908-0920-003/doc/94327601. Accessed October 1, 2021.
20. Hinde J, Demétrio CGB. Overdispersion: models and estimation. *Comput Stat Data Anal.* 1998;27(2):151–170.
21. Hinde J. Compound Poisson regression models. In: Gilchrist R, ed. *GLIM 82: Proceedings of the International Conference on Generalised Linear Models*. New York, NY: Springer-Verlag New York; 1982:109–121.
22. Getis A. A history of the concept of spatial autocorrelation: a geographer's perspective. *Geogr Anal.* 2008;40(3):297–309.
23. Liu B, Zhu L, Zou J, et al. Updated methodology for projecting U.S.- and state-level cancer counts for the current calendar year: part I: spatio-temporal modeling for cancer incidence. *Cancer Epidemiol Biomarkers Prev.* 2021;30(9):1620–1626.
24. Besag J, York J, Mollié A. Bayesian image restoration, with two applications in spatial statistics. *Ann Inst Stat Math.* 1991;43(1):1–20.
25. Banerjee S. Revisiting spherical trigonometry with orthogonal projectors. *Coll Math J.* 2004;35(5):375–381.
26. Morales-Otero M, Núñez-Antón V. Comparing Bayesian spatial conditional overdispersion and the Besag–York–Mollié models: application to infant mortality rates. *Mathematics.* 2021;9(3):282.
27. Arab A. Spatial and spatio-temporal models for modeling epidemiological data with excess zeros. *Int J Environ Res Public Health.* 2015;12(9):10536–10548.
28. Jay M, Oleson J, Charlton M, et al. A Bayesian approach for estimating age-adjusted rates for low-prevalence diseases over space and time. *Stat Med.* 2021;40(12):2922–2938.
29. Asmarian N, Ayatollahi SMT, Sharafi Z, et al. Bayesian spatial joint model for disease mapping of zero-inflated data with R-INLA: a simulation study and an application to male breast cancer in Iran. *Int J Environ Res Public Health.* 2019;16(22):E4460.
30. Lunn D, Spiegelhalter D, Thomas A, et al. The BUGS project: evolution, critique and future directions. *Stat Med.* 2009;28(25):3049–3067.
31. Lindgren F, Rue H. Bayesian spatial modelling with R-INLA. *J Stat Soft.* 2015;63(19):1–25.
32. Blangiardo M, Cameletti M, Baio G, et al. Spatial and spatio-temporal models with R-INLA. *Spat Spatio-temporal Epidemiol.* 2013;4:33–49.
33. Gerber F, Furrer R. Pitfalls in the implementation of Bayesian hierarchical modeling of areal count data: an illustration using BYM and Leroux models. *J Stat Soft Code Snippets*. 2015;63(1):1–32.
34. Rue H, Riebler A, Sørbye SH, et al. Bayesian computing with INLA: a review. *Annu Rev Stat Appl.* 2017;4(1):395–421.
35. Gardner W, Mulvey EP, Shaw EC. Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models. *Psychol Bull.* 1995;118(3):392–404.
36. Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Dent Tech.* 1992;34(1):1.
37. Mullahy J. Specification and testing of some modified count data models. *J Econom.* 1986;33(3):341–365.
38. Heilbron DC. Zero-altered and other regression models for count data with added zeros. *Biom J.* 1994;36(5):531–547.
39. Hankey BF, Ries LA, Edwards BK. The Surveillance, Epidemiology, and End Results program: a national resource. *Cancer Epidemiol Biomarkers Prev.* 1999;8(12):1117–1121.
40. Pickle LW, Feuer EJ, Edwards BK. Prediction of incident cancer cases in non-SEER counties. In: *Proceedings of the Biometrics Section of the 2000 Annual Meeting of the American Statistical Association*. Alexandria, VA: American Statistical Association; 2001:45–52.
41. Mungiole M, Pickle LW, Simonson KH. Application of a weighted head-banging algorithm to mortality data maps. *Stat Med.* 1999;18(23):3201–3209.
42. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. *CA Cancer J Clin.* 2020;70(1):7–30.
43. Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* 2021;71(3):209–249.