An R package for the integrative evaluation of statistical approaches for cancer incidence projections

Maximilian Knoll (m.knoll@dkfz.de)
Deutsches Krebsforschungszentrum https://orcid.org/0000-0002-9037-3980
Jennifer Furkel
Deutsches Krebsforschungszentrum
Jürgen Debus
UniversitatsKlinikum Heidelberg
Amir Abdollahi
Deutsches Krebsforschungszentrum
André Karch
Westfalische Wilhelms-Universitat Munster
Christian Stock
Universitatsklinikum Heidelberg Institut fur Medizinische Biometrie und Informatik

Research article

Keywords: Age-Period-Cohort model, Bayesian model, Cancer incidence projection, INLA, Rate projection

Posted Date: June 12th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-34369/v1

Version of Record: A version of this preprint was published on October 15th, 2020. See the published version at https://doi.org/10.1186/s12874-020-01133-5.

Abstract

Background Projection of future cancer incidence is of high importance for epidemiological and biomedical research. Age-Period-Cohort (APC) models are established for such projections, usually based on long-term cancer registry data (>20yrs). In many countries (including Germany), however, nationwide long-term data are not yet available. It is unclear which statistical approach should be recommended for projections using rather short-term data. Methods An R package (incAnalysis) for the comparative analysis of projection methods was developed, supporting in particular Bayesian models fitted by Integrated Nested Laplace Approximations (INLA). It was used to assess operating characteristics (bias, coverage, precision) of various statistical approaches to cancer incidence projection based on artificially shortened long-term data from three cancer registries (SEER-9, NORDCAN, Saarland). Additionally, a novel age-period INLA z-model with multivariate tensor product smoothers is described. Results Coverage was high (mostly >90%) for Bayesian APC models (BAPC), whereas less complex models showed differences in coverage dependent on projection-period. Intercept-only models yielded values below 20% for coverage. Bias increased and precision decreased for longer projection periods for all except intercept-only models. Precision was lowest for BAPC, multivariate smoother and Age-Period GLMs with interaction effect. Conclusion The incAnalysis R package allows for straightforward comparison of (cancer incidence) rate projection approaches. Of the evaluated model classes, more complex models showed an increase in coverage at the cost of precision; bias increased for longer projection intervals, mostly irrespective of the level of complexity.

Background

Evaluation of cancer prevention and screening programs require reliable estimates of future cancer incidence [1, 2]. Projections are often performed using long-term data (>20yrs) from population-based cancer registries [3]. For short-term data, it is less clear and there appears to be a lack of guidance which statistical approach is to be recommended. The need to base projection models on relatively short-term data is relevant e.g. for Germany, where aggregated data of cancer incidence on a national level is only available from 1999 on, as well as for many countries with newly established cancer registries.

A selection of previously applied projection models is outlined in [4]. Relatively simple approaches assuming constant rates were utilized [5, 6], as well as more complex age-period (AP) models formulated as generalized linear models (GLMs) [7–9]. Clements et al. use generalized additive models (GAMs) [10]. GAMs can include uni- or multivariate smoothers in their linear predictors. An established model class for incidence projections based on long-term observation data are age-period-cohort (APC) models, which additionally incorporate a cohort effect [11, 12]. Even though projections of APC usually yield robust results, the APC identification problem impairs direct interpretability of single effects [13, 14].

Projection models are often fitted within a classical maximum likelihood (ML) or restricted maximum likelihood (REML) framework [15–17]. Alternatively, a Bayesian framework may be used [18, 19]. Bayesian model estimation can be implemented using Markov-Chain Monte Carlo (MCMC) methods,

which show the disadvantage of being computationally intensive. A recently developed computationally far less demanding alternative are Integrated Nested Laplace Approximations (INLA) [20, 21].

GAMs usually incorporate splines to fit univariate trends or tensor product smoothers for multivariate trends (i.e. interactions between function of continuous variables). In the classical frequentist framework, such models can be fit e.g. using the mgcv-package in R [28]. In the latter, uni- and multivariate smoothers can directly be incorporated in the model formula, e.g. as splines or tensor product smoothers.

Recently, a highly flexible Bayesian APC (BAPC) model based on the INLA approach has been proposed for future cancer incidence projections which assumes a Poisson distribution of incidence counts [19]. Havulinna et al. demonstrate that interactions between effects can be modeled by specifying appropriate priors [18].

Given the lack of guidance on statistical modeling approaches for cancer incidence projection, we evaluated a range of possibly suitable models using a dedicated R package which allows a direct, integrative comparison of model performance metrics. Due to the desirable flexible modeling options and the probabilistic interpretation of results in a Bayesian framework as well as the computationally very efficient implementation, we emphasize the INLA approach. We use artificially shortened parts of observed long-term data from three population-based cancer registries to empirically evaluate predictive performance (bias, coverage and precision) of various statistical models, starting with simple models, e.g. intercept-only models and increasing complexity to BAPC models. Further, we present a novel approach to directly implement multivariate tensor product smoothers for age and period as a *z*-model in INLA.

Methods

Cancer registry data

Three low incident tumor sites/entities (brain tumors, kidney cancer, melanoma) and four high incident entities (lung, breast, colorectal, prostate) were selected from three population-based cancer registries: SEER–9 [22], NORDCAN [23] and Saarland [24]. Specific selection criteria are shown in Table 1.

From the Surveillance, Epidemiology, and End Results (SEER) Program in the United States, SEER–9 cancer incidence data (1973–2014) were accompanied by population data, available in one year age groups.

NORDCAN data, comprise cancer incidence data from Denmark, Finland, Iceland, Norway, Sweden, Faroe Islands and Greenland were retrieved from the NORDCAN website on 2018–08–01. Incidence data were available in five years groups. Population matrices were calculated from the person-years at risk information.

Cancer incidence data from Saarland, a German federal state with a long-established cancer registry, were obtained from the Saarland cancer registry website on 2018–08–01 (5yr age groups). Population data were retrieved from the health report system of the federal government (up to 2012) und from the website of Saarland for the years 2013/14 [25, 26].

Table 1: Selection details for analyzed tumor sites/entities for the three cancer registries. $^{-}$low, $^{+}$high incidence.

registry	**entity/site**	**selection**
SEER-9		
	Glioblastoma[-]	HISTO3V: 9440
	Kidney cancer[-]	PRIMSITE: C649
	Melanoma[-]	PRIMSITE: C440-449
	Lung and bronchial tumors[+]	PRIMSITE: C340-349
	Breast cancer[+]	PRIMSITE: C500-509
	Colorectal cancer[+]	PRIMSITE: C18-20
	Prostate cancer[+]	PRIMSITE: C61.9
NORDCAN		
	Brain, central nervous system[-]	cancer: 340
	Kidney[-]	cancer: 290
	Melanoma of skin[-]	cancer: 310
	Lung[+]	cancer: 180
	Breast[+]	cancer: 200
	Colorectal[+]	cancer :590
	Prostate[+]	cancer:261
Saarland		
	Brain tumors [Gehirn] [-]	loc: 191
	Kidney cancer [Niere, sonst.u.n.n.bez. Harnorgane] [-]	loc: 189
	Melanoma [Bösartiges Melanom der Haut] [-]	loc: 712
	Lung, bronchial and tracheal tumors [Luftröhre, Bronchien u. Lunge] [+]	loc: 162
	Female breast tumors [Weibliche Brustdrüse] [+]	loc: 174
	Colorectal cancer [Dick- und Mastdarm] [+]	loc: 153+154

Prostate cancer [Prostata] +	loc: 185

Projection models

The incAnalysis R package (see details below in section 3.2) was used to evaluate a number of increasingly complex models (GLMs, GAMs, BAPC) using the INLA framework. To describe the evaluated models, we introduce the following notation: *Y* denotes observed cancer incidence counts, *N* denotes population size, *AGE* and *PERIOD* are the respective covariates. The notation also corresponds to variable names used in the R package. Age or age-group, respectively, is indexed by *i.*

GLMs are formulated using three components: (1) a probability distribution from the exponential family, (2) a linear predictor η = Xβ and (3) a link function *g* with Ey = μ = g−1(η). In all, except BAPC models, negative-binomially distributed counts of tumor cases were assumed.

The most simplistically structured GLM includes only an intercept, η = β0. In R, this intercept-only model was formulated as Y ~ offset(log(N)).

Next, a GLM with age and period as covariates together with their interaction term was assessed: η = β0+ β1age+ β2period+ β3age:period, corresponding to the R formula Y~offset(log(N))+ AGE*PERIOD.

GAMs have a structure similar to GLMs, with the difference that smooth functions fis of covariates can be included in the linear predictor (*A:* model matrix, θ: parameter vector): gμ = A θ+f1x1+f2x2+….

Splines might be used as smooth functions, or in the case of INLA, specific Gaussian Markov Random Fields. In the present analysis, B-splines were used as univariate smoother for the age covariate and bs() from the splines package can directly be included in the model formula: Y~offset(log(N))+PERIOD+bs(AGE). Alternatively, an rw2 model might be specified as Y~offset(log(N))+PERIOD+f(AGE, model = ‘rw2’).

Next, a multivariate smoother for age and period was evaluated. For this purpose, tensor spline interactions can be specified, e.g. by using the function mgcv::te() for the classical model fitting approach (Y~offset(log(N))+te(AGE,PERIOD)).

In INLA models, te() is not directly usable in model formulas. We propose to use a *z*-model instead, which is an implementation of classical random effects part of a mixed model (η = …+Z z).

Z was calculated as the tensor product smooth model matrix for marginal bases for age and period using mgcv::tensor.prod.model.matrix(). Marginal bases were calculated as M-splines, using splines2::mSpline(). M-splines are non-negative splines, which can be considered as a normalized version

of B-splines. A loggamma prior was specified for this model, with param values of (1, 0.005), the same values used as in [27]. The corresponding R code is shown in the package vignette vignette('incidence').

Performance metrics

Model performance was evaluated using three metrics: coverage, bias and precision. Metrics were calculated per age/age-group, sex and entity, and averaged.

Coverage was calculated as the fraction of projections laying within the 95% (equal tailed) credibility band. Bias was set to 0 if the observed incidence count was equal to the predicted, otherwise the ratio (observed-predicted)/observed was computed. INLA reported posterior standard deviations were used to measure precision.

Model performance

Evaluation of the predictive performance of models with increasing complexity was performed as follows (see also Fig. 1): the most current observed incidence data was predicted, with the projection period starting *n* years prior to this timepoint ($n \in \{2,5,10,15,20\}$). The observation period for model training preceded this timepoint. In the presented analysis, 15yrs were chosen as observation period. For the evaluation of a 2yr projection, e.g. in the SEER–9 dataset, data of the year 2014 would be predicted, using data from the 15yrs prior to 2012 for model fitting.

Das was available in different aggregation types - as age-groups for NORDAN and Saarland data and for each age for the SEER–9 data. In the latter case, individual age-years were used, i.e. no further aggregation was applied.

R package incAnalysis

To facilitate analysis, reproducibility and further applications, the R package 'incAnalysis' was developed. It is publicly available on http://github.com/mknoll/incAnalysis. The package mainly builds on methods in the R packages BAPC [19] and mgcv [28]. Representative analyses with stepwise explanations on how to use the package are outlined in the accompanying vignette in more detail: vignette('incidence')in R. An overview of the functionality and structure of the package is given in Fig. 2.

A wide variety of approaches to project future cancer incidence can be comparatively tested using this package. Constant rates or counts simply projected into the future, as well as GLMs and GAMs (both in the INLA and ML/REML framework, selected via the method parameter) and BAPC models might be specified.

The package provides a class called incClass which is instantiated with population and incidence data (data.frame with years in rows, the earliest available year in the first row and age/age-group as columns with increasing values from left to right) as well as the period used for model training and the fitting period of interest (and additional parameters). Different models are then added to the newly created object with the following functions which usually expect additional parameters, e.g. model formulas and the respective class object: runFwProj() for forward projection of constant rates or constant counts, runGLM() for generalized linear models (using INLA or an ML approach, specified by the method parameter), runGAM() for GAMs, runInla() for any INLA model and runBAPC() to run the BAPC model [19]. evaluate() calculates the performance metrics, which can be extracted as data.frame via metrics(); additionally, projections are plotted. pitHist() plots Probability Integral Transform (PIT) histograms for all INLA fitted models.

Results

Coverage

Coverages for the evaluated models are shown in Fig. 3 for an observation period of 15yrs and projection periods of 2, 5, 10, 15 and 20yrs.

Importantly, most models yielded coverages below 95%, with smallest (<25%) coverages for intercept only models and highest coverages (>75%) for BAPC models, irrespective of the projection period. Variability of coverages of BAPC projections is smaller in the SEER–9 dataset as compared to NORDCAN and Saarland data.

Coverage increased for AP models with linear age, period and interaction effect for longer projection intervals in all datasets. Models incorporating a univariate smoother for age showed no clear median increase in coverages for longer periods, variability, however, increased.

Multivariate smoother models showed a decrease of median coverages for longer projection intervals in the SEER–9 data, in increase in the Saarland data and high variability with no clear trend in the NORDCAN data.

Bias

Results of bias analyses are shown in Fig. 4. Negative values correspond to higher predicted than observed incidence counts (overestimation). For visualization purposes, values <–200 were set to –200.

Several models show negative values. Absolute bias increases with longer projection intervals for most models in the SEER–9 and Saarland datasets. Intercept-only models show mostly absolute median bias values below –100, except for 15 and 20yr projections in the Saarland data. Univariate smoother models show in most cases lower absolute bias as GLMs with linear age, period and interaction effects. Median

absolute bias is smallest for the multivariate smoother models in SEER–9 data for longer projection intervals. Differences in median absolute bias between all except intercept-only models are highest in the SEER–9 dataset.

Precision

Precision is depicted in Fig. 5; median model values range mostly between 0.5 and 5 for the SEER–9 data, 2 and 6 for the NORDCAN data and 0 and 4 in the Saarland dataset. Longer projection intervals yield lower precision for all but the intercept only model. Univariate smoother models show higher precision as compared to most additionally evaluated models. Variability in precision increases for longer projection intervals for the BAPC models, and for the SEER–9 data, for univariate smoother GAMs. For the other models, no clear trend can be observed.

Discussion

Population-based cancer registry data are routinely used to monitor cancer incidence at the population-level, to evaluated screening and prevention programs, and to identify areas where intensified medical research is needed [4]. However, no recommendations exists on which models to use for projections based on short-term observational rate data.

We developed an R package for the integrative evaluation of different model types for rate projections and utilized it for the systematic evaluation of historic cancer incidence data from multiple cancer registries. This allowed to obtain recommendations about projection intervals and model types. Importantly, any kind of rate data as e.g. mortality data, can be evaluated analogously.

Only age(-groups) between 20 and 84 were analyzed, as childhood tumors constitute a biologically distinct group and are in general rare. This might impair the ability of models to obtain reliable projections; nevertheless it has been reported [29] that this approach might decrease accuracy. Cancers in the age group > = 85 were excluded to assure comparability between cancer registries.

Model performance was assessed by evaluating coverage, bias and precision of projections. To ease calculations, the R package incAnalysis was developed. Alternative metrics for model evaluation described are e.g. the Continuous Ranked Probability Score (CPRS) as used e.g. in [19] or the evaluation of PIT histograms. The latter can be easily obtained from INLA fitted objects, and further metrics as the CPRS can be easily calculated using the data provided by the incAnalysis package.

As least complex model, intercept only models were evaluated. As expected, only small coverages (<25%) could be expected as cancer occurrence is usually highly dependent on age. An intercept only model does not take the age into accounts, and thus, these models cannot be recommended for cancer incidence projection.

GLMs with linear age, period and their interaction effect were evaluated as next, more complex model types. Performance, however, was generally poor. To achieve a potentially even better fit, a model with a univariate smoother for age was analyzed, as the latter is a biologically highly relevant covariate for cancer incidence. B-splines, created with splines::bs() were incorporated into the model formula. An alternative would be the specification of a Gaussian Markov Random Field structure for smoothing, e.g. a second order random walk.

Next, multivariate smoothers (tensor product smoothers) for age and period were included into the model, using a newly proposed *z*-model in INLA. For classical ML/REML models, such effects can easily be included in the models by using the mgcv::te() function. The latter cannot be directly fit with INLA::inla(). Even though the mgcv::ginla() function was made available recently (which allows to obtain posterior distributions of effects directly from GAMs fitted with mgcv), the INLA package is not directly utilized by mgcv, and thus projections are not as straight-forward as with the newly proposed *z*-model. Coverage is higher as compared to univariate smoother models, but is less stable for long term projections as compared to BAPC models.

Finally, the BAPC model was evaluated which assumes a Poisson distributions and includes the three random effects age, period, cohort (rw2) and an additional random effect (iid) to adjust for overdispersion. Models were located among the best performing for all evaluated parameter combinations. The additional two effects (cohort and overdispersion adjustment effect) seem to be especially important for short-term projections, as differences to most other models except multivariate smoother models decrease for longer intervals.

Conclusions

Projections of rate data using short term data yields robust high coverage at the cost of low precision for BAPC. Less complex models mostly yield better results for longer projection intervals (>10 yrs). Tensor product smooth models (age, period) constitute a reasonable alternative. Intercept only models should only be used for short projections (<5 yrs). An easy-to-use R package was developed for the present analysis which allows comparative evaluation of any kind of rate data and is publicly available.

List Of Abbreviations

APC Age-Period-Cohort

BAPC Bayesian APC models

CPRS Continuous Ranked Probability Score

GAM Generalized Additive Model

GLM Generalized Linear Model

INLA Integrated Nested Laplace Approximations

MCMC Markov-Chain Monte Carlo

ML Maximum Likelihood

PIT Probability Integral Transform

REML Restricted Maximum Likelihood

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The datasets generated and/or analysed during the current study are included in the incAnalysis github package, https://github.com/mknoll/incAnalysis.

Competing interests

CS is now full-time employee of Boehringer Ingelheim Pharma GmbH & Co. KG, Ingelheim, Germany. The company had no role in design, analysis or interpretation of the presented work.

Funding

Nationales Centrum für Tumorerkrankungen Heidelberg (NCT PRO–2015.21), Deutsche Forschungsgemeinschaft (DFG UNITE SFB–1389), Deutsches Krebsforschungszentrum (iMed). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Authors' contributions

CS designed the study, MK designed and created the R package. MK and CD wrote the manuscript with input from all authors. All authors provided critical feedback and helped shape the research, analysis and manuscript.

Acknowledgements

MK and JF are members of the MD/PhD program at Heidelberg University and are funded by Heidelberg Medical Faculty.

References

1.Brown LD, Cai TT, DasGupta A, Agresti A, Coull BA, Casella G, Corcoran C, Mehta C, Ghosh M, Santner TJ *et al: Interval estimation for a binomial proportion - Comment - Rejoinder. Statistical Science* 2001, *16(*2):101–133.

2.Siegel RL, Miller KD, Jemal A: *Cancer statistics, 2019. CA Cancer J Clin* 2019, *69(*1):7–34.

3.Moller B, Fekjaer H, Hakulinen T, Sigvaldason H, Storm HH, Talback M, Haldorsen T: *Prediction of cancer incidence in the Nordic countries: empirical comparison of different approaches. Stat Med* 2003, *22(*17):2751–2766.

4.Bray F, Moller B: *Predicting the future burden of cancer. Nat Rev Cancer* 2006, *6(*1):63–74.

5.Moller H, Fairley L, Coupland V, Okello C, Green M, Forman D, Moller B, Bray F: *The future burden of cancer in England: incidence and numbers of new patients in 2020. Br J Cancer* 2007, *96(*9):1484–1488.

6.Nowatzki J, Moller B, Demers A: *Projection of future cancer incidence and new cancer cases in Manitoba, 2006–2025. Chronic Dis Can* 2011, *31(*2):71–78.

7.Dyba T, Hakulinen T, Paivarinta L: *A simple non-linear model in incidence prediction. Stat Med* 1997, *16(*20):2297–2309.

8.Hakulinen T, Dyba T: *Precision of incidence predictions based on Poisson distributed observations. Stat Med* 1994, *13(*15):1513–1523.

9.Stock C, Mons U, Brenner H: *Projection of cancer incidence rates and case numbers until 2030: A probabilistic approach applied to German cancer registry data (1999–2013). Cancer Epidemiol* 2018, *57:*110–119.

10.Clements MS, Armstrong BK, Moolgavkar SH: *Lung cancer rate predictions using generalized additive models. Biostatistics* 2005, *6(*4):576–589.

11.Engeland A, Haldorsen T, Tretli S, Hakulinen T, Horte LG, Luostarinen T, Schou G, Sigvaldason H, Storm HH, Tulinius H *et al: Prediction of cancer mortality in the Nordic countries up to the years 2000 and 2010, on the basis of relative survival analysis. A collaborative study of the five Nordic Cancer Registries. APMIS Suppl* 1995, *49:*1–161.

12.Smith TR, Wakefield J: *A Review and Comparison of Age-Period-Cohort Models for Cancer Incidence. Statistical Science* 2016, *31(*4):591–610.

13.Kupper LL, Janis JM, Salama IA, Yoshizawa CN, Greenberg BG. *Age-Period-Cohort Analysis - an Illustration of the Problems in Assessing Interaction in One Observation Per Cell Data, Commun Stat-Theor M* 12(23) (1983) 2779–2807.

14.O'Brien RM, *Constrained Estimators and Age-Period-Cohort Models, Sociological Methods & Research* 40(3) (2011) 419–452.

15.Mistry M, Parkin DM, Ahmad AS, Sasieni P: *Cancer incidence in the United Kingdom: projections to the year 2030. Br J Cancer* 2011, *105(*11):1795–1803.

16.Moller B, Fekjaer H, Hakulinen T, Tryggvadottir L, Storm HH, Talback M, Haldorsen T: *Prediction of cancer incidence in the Nordic countries up to the year 2020. Eur J Cancer Prev* 2002, *11 Suppl 1:*S1–96.

17.Whiteman DC, Green AC, Olsen CM: *The Growing Burden of Invasive Melanoma: Projections of Incidence Rates and Numbers of New Cases in Six Susceptible Populations through 2031. J Invest Dermatol* 2016, *136(*6):1161–1171.

18.Havulinna AS: *Bayesian age-period-cohort models with versatile interactions and long-term predictions: mortality and population in Finland 1878–2050. Stat Med* 2014, *33(*5):845–856.

19.Riebler A, Held L: *Projecting the future burden of cancer: Bayesian age-period-cohort analysis with integrated nested Laplace approximations. Biom J* 2017, *59(*3):531–549.

20.Rue H, Martino S, Chopin N: *Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. JSTOR* 2009, *71(*2):319–392.

21.Rue H, Riebler Am Sorbye, SH, Illian JB, Simpson DP, Lindgren, *FK. Bayesian Computing with INLA: A Review. Annu Rev Stat Appl* 4 (2017) 395–421.

22.*Surveillance Epidemiology and End Results (SEER) Program* (www.seer.cancer.gov),

Research Data (1973–2014), *National Cancer Institute,* DCCPS, Surveillance Research Program, based on the November 2016 submission., 2016. *https://seer.cancer.gov.*

23.Engholm G, Ferlay J, Christensen N, Bray F, Gjerstorff M, Klint A, Kotlum J, Olafsdotti E, Pukkala E, Storm H. *NORDCAN—a Nordic tool for cancer information, planning, quality control and research, Acta*

Oncol 49(5) (2010) 725–36.

24.Krebsregister Saarland. *http://www.krebsregister.saarland.de/.* Accessed 2018–10–25.

25.Das Statistische Amt des Saarlandes, Tabellen und Grafiken aus dem Bereich “Gebiet und Bevölkerung”, 2018. *https://www.saarland.de/6772.htm.* Accessed 2018–10–25.

26.Gesundheitsberichterstattung des Bundes, Bevölkerung im Jahresdurchschnitt 1980–2012 (Grundlage Zensus BRD 1987, DDR 1990), 2018. *http://www.gbe-bund.de/gbe10/trecherche.prc_them_rech?tk = 700&tk2 = 906&p_uid = gast&p_aid = 66019368&p_sprache = D&cnt_ut = 1&ut = 906.* Accessed 2018–10–25.

27.Bauer C, Wakefield J, Rue H, Self S, Feng ZJ, Wang Y: *Bayesian penalized spline models for the analysis of spatio-temporal count data. Statistics in Medicine* 2016, *35(*11):1848–1865.

28.Wood SN, *Generalized Additive Models: An Introduction with R,* Second Edition ed., *Chapman and Hall/CRC Texts in Statistical Science,* Boca Raton, 2017.

29.Baker A, Bray I: *Bayesian projections: what are the effects of excluding data from younger age groups? Am J Epidemiol* 2005, *162(*8):798–805.

Figures

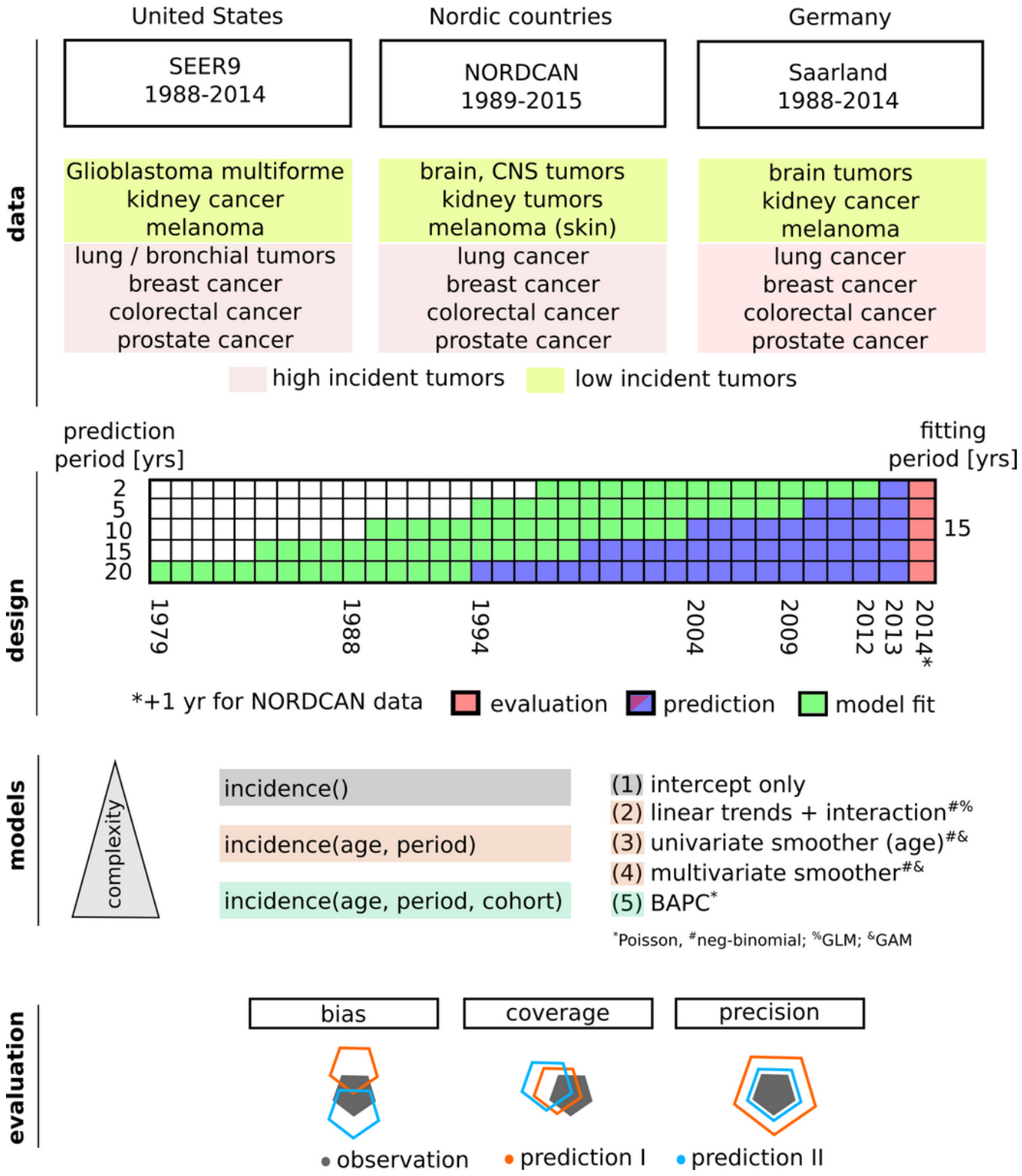

Figure 1

Overview of the analyzed cancer registry data, study design, model selection and evaluation metrics.

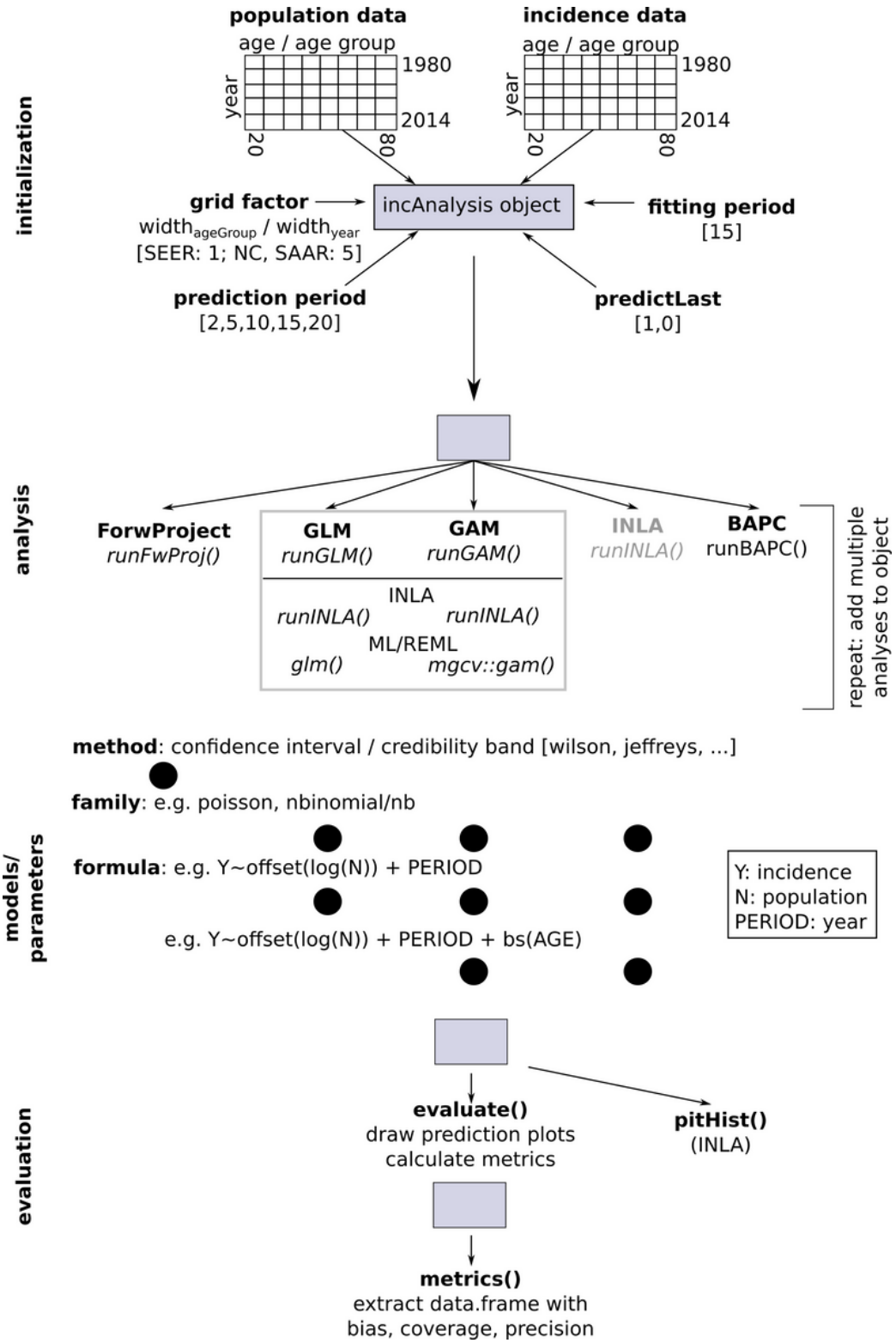

Figure 2

The R incAnalysis package.

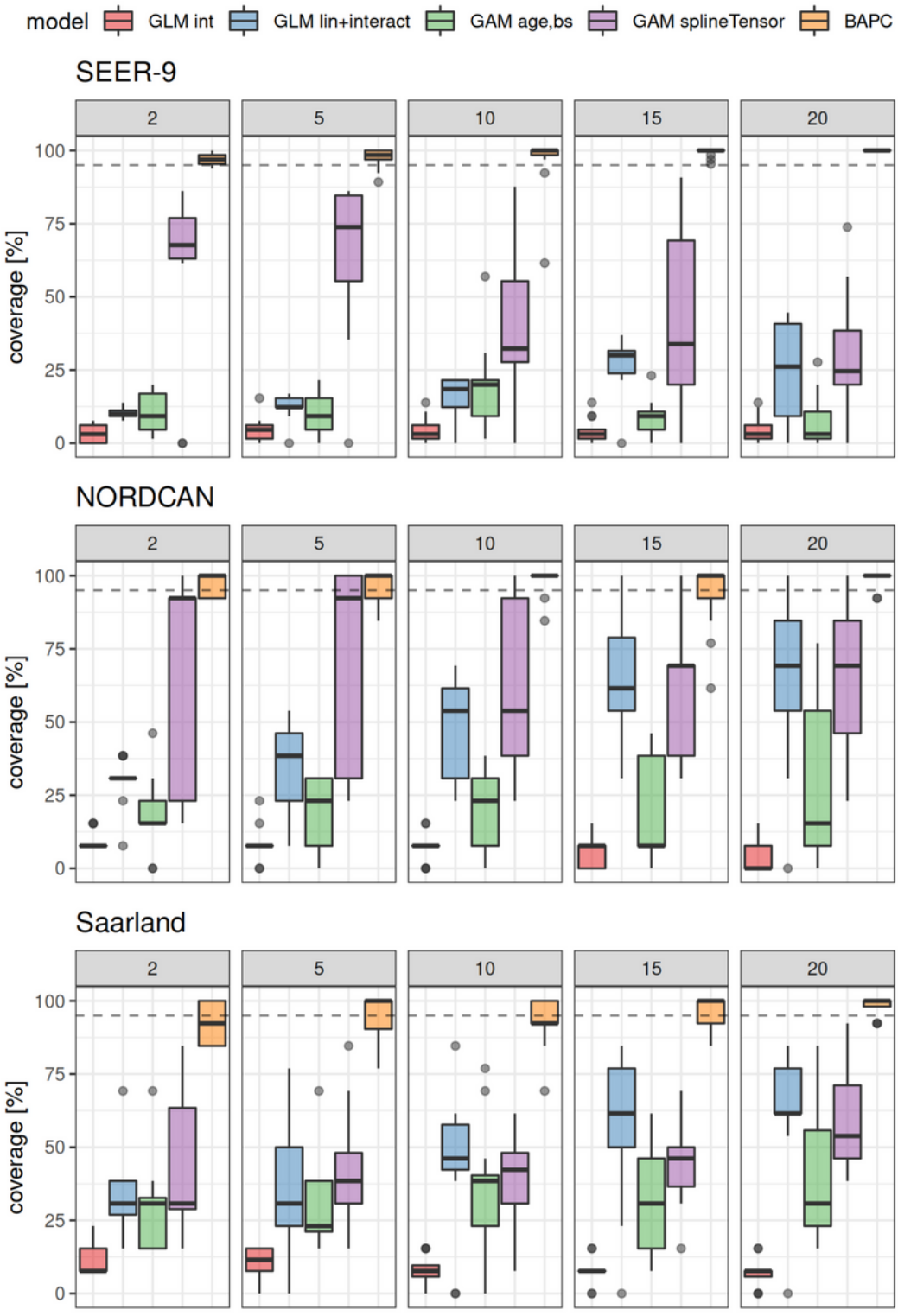

Figure 3

Coverages of future projections after 2, 5, 10, 15 and 20yrs based on models with a 15yr observation period. Dashed line: 95% coverage. int: intercept only model, lin+interact: linear age, period and interaction effects, age,bs: univariate smoother (B-spline) for age, splineTensor: tensor product smoother (age, period), M-spline basis. GLMs, GAMs: neg-binomial distribution.

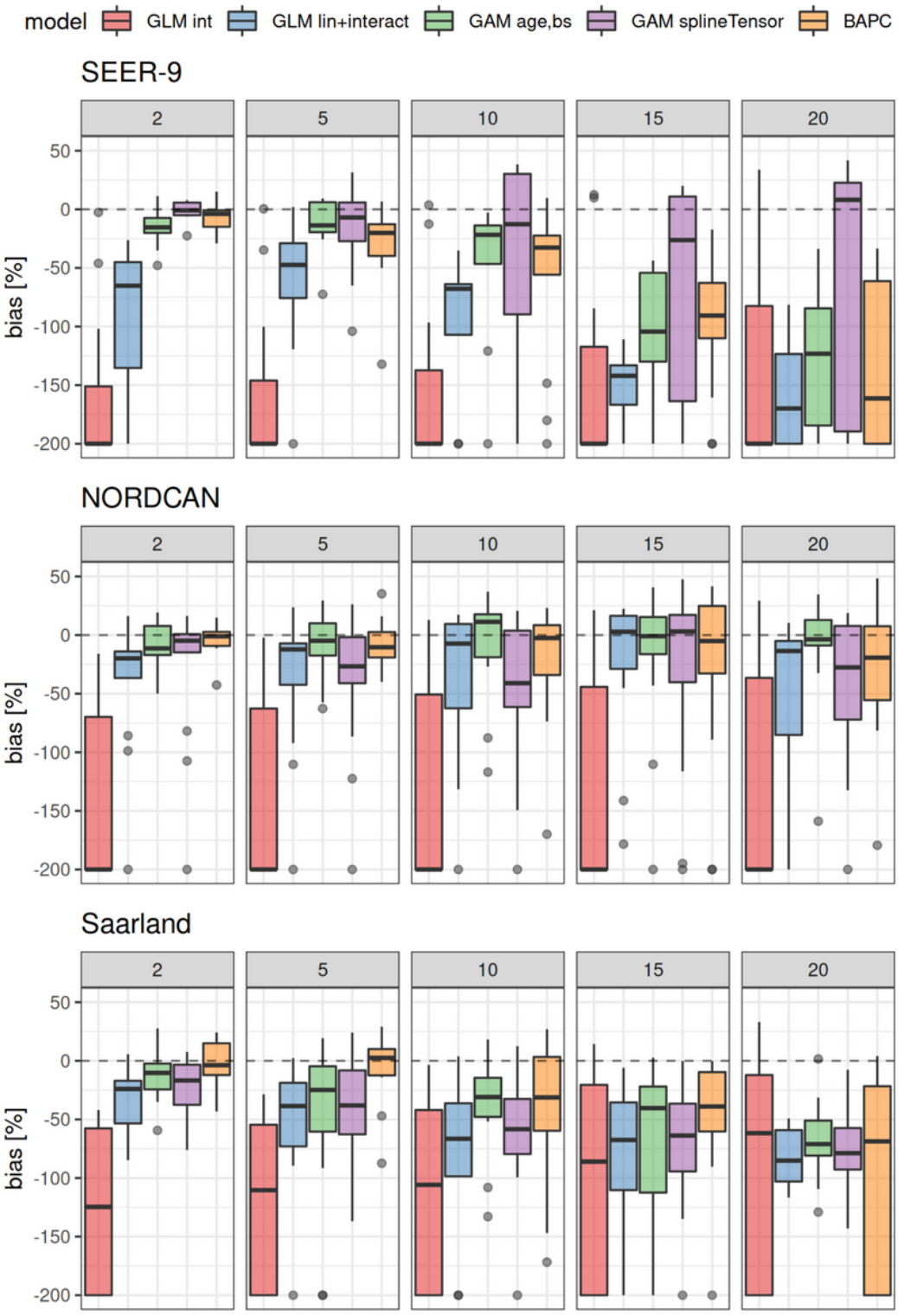

Figure 4

Bias of future projections after 2, 5, 10, 15 and 20yrs based on models with a 15yr observation period. Negative values indicate overestimation of cancer incidence. Bias values smaller than -200 were set to -200. Dashed line: no bias (0%). int: intercept only model, lin+interact: linear age, period and interaction effects, age,bs: univariate smoother (B-spline) for age, splineTensor}: tensor product smoother (age, period), M-spline basis. GLMs, GAMs: neg-binomial distribution.

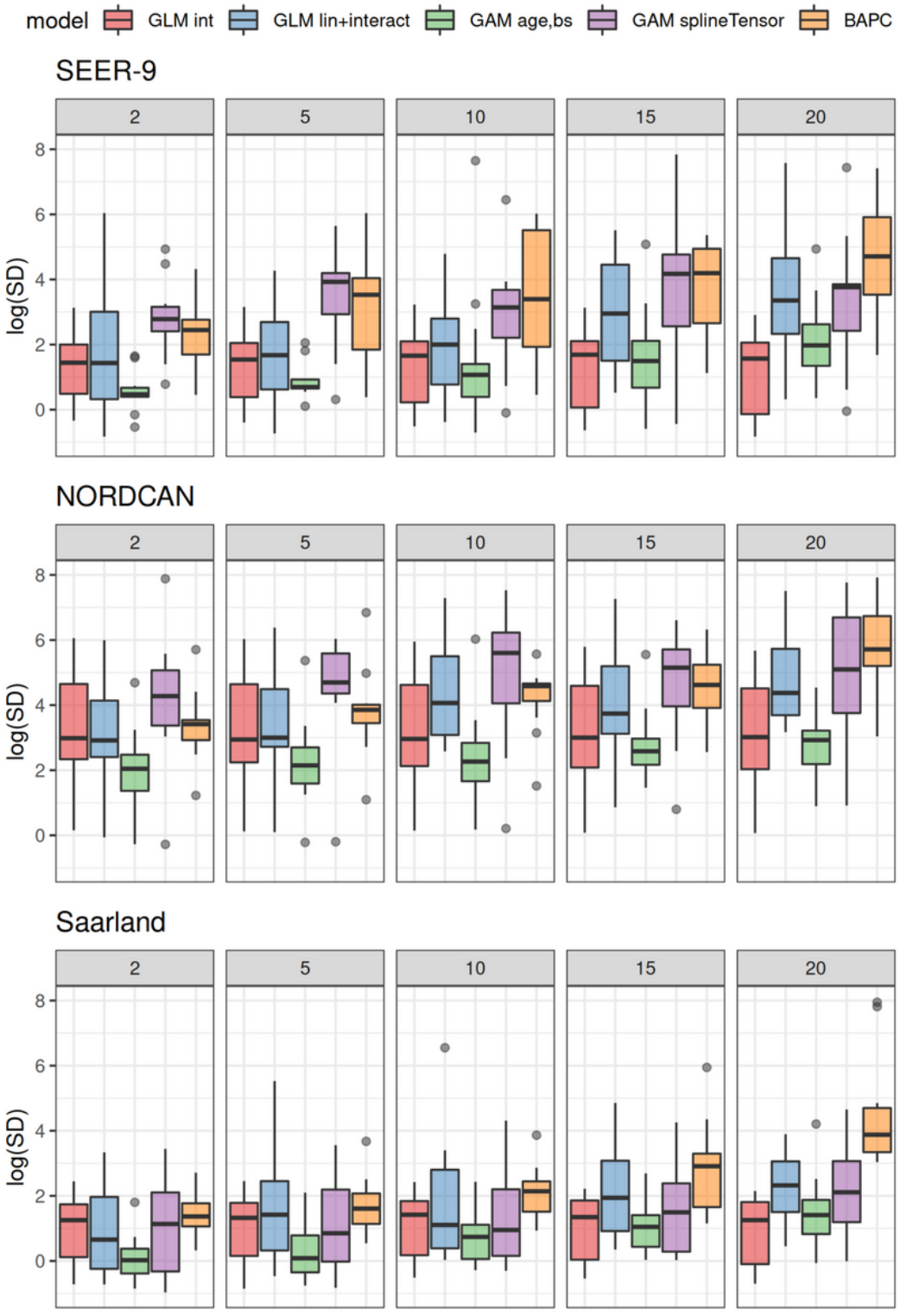

Figure 5

Precision of future projections after 2, 5, 10, 15 and 20yrs based on models with a 15yr observation period. Transformed averaged posterior standard deviations are shown. int: intercept only model, lin+interac: linear age, period and interaction effects, age,bs: univariate smoother (B-spline) for age, splineTensor: tensor product smoother (age, period), M-spline basis. GLMs, GAMs: neg-binomial distribution.