

## ORIGINAL ARTICLE

# Personalised need of care in an ageing society: The making of a prediction tool based on register data

Marvin N. Wright<sup>1</sup> | Sasmita Kusumastuti<sup>2</sup> | Laust H. Mortensen<sup>2,3</sup> |  
Rudi G. J. Westendorp<sup>2</sup> | Thomas A. Gerds<sup>4</sup>

<sup>1</sup>Leibniz Institute for Prevention Research & Epidemiology–BIPS, Bremen, Germany

<sup>2</sup>Section of Epidemiology, Department of Public Health, University of Copenhagen, Copenhagen, Denmark

<sup>3</sup>Section of Methodology and Analysis, Statistics Denmark, Copenhagen, Denmark

<sup>4</sup>Section of Biostatistics, Department of Public Health, University of Copenhagen, Copenhagen, Denmark

## Correspondence

Marvin N. Wright, Leibniz Institute for Prevention Research & Epidemiology–BIPS, Achterstr. 30, 28359 Bremen, Germany.

Email: [wright@leibniz-bips.de](mailto:wright@leibniz-bips.de)

## Funding information

Novo Nordisk Foundation Challenge Programme, Grant/Award Number: NNF17OC0027812

## Abstract

Danish municipalities monitor older persons who are at high risk of declining health and would later need home care services. However, there is no established strategy yet on how to accurately identify those who are at high risk. Therefore, there is great potential to optimise the municipalities' prevention strategies. Denmark's comprehensive set of electronic population registers provide longitudinal data that cover individual and household socio-demographics and medical history. Using these data, we developed and applied recurrent neural networks to predict the risk of a need of care services in the future and thus identify individuals who would benefit the most from the municipalities' prevention strategies. We compared our recurrent neural network model to prediction models based on Cox regression and Fine–Gray regression in terms of calibration and discrimination. Challenges for the prediction modelling were the competing risk of death and the longitudinal information on the registered life course data.

## KEYWORDS

competing risks, deep learning, register data, recurrent neural networks, survival analysis

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2021 The Authors. Journal of the Royal Statistical Society: Series A (Statistics in Society) published by John Wiley & Sons Ltd on behalf of Royal Statistical Society

# 1 | INTRODUCTION

Providing optimal and resource-efficient care in an ageing society requires preventive interventions that are targeted towards older persons at high risk of declining health. In Denmark, anyone can ask for home care services from the municipality and the municipality has to ensure that everyone gets care when needed. As of 1 July 2019, Danish municipalities by law must assess the need of home care for all residents who turn 70 years old and living alone, or turn 75 and 80 regardless of their living situation. For everyone else between the ages of 65 and 80 years, the municipality must monitor which older persons are at high risk of declining health and in need of home care services (Borger.dk, 2019). However, so far there is no established strategy on how to accurately identify these older persons at high risk.

The overarching aim of this project is to build an algorithm that can identify older persons who are at high risk of declining health that would require long-term personal home care services. Here, we used a comprehensive set of register data on health and socio-demographics and developed methods to predict the future use of municipal home care services. Eventually, our models will be implemented to provide person-specific information in the form of an online tool to assist municipalities with the information on which older persons are at greater need of a preventive efforts than others. This online tool will help to optimise the municipalities' preventive strategies for older citizens to delay functional decline and diminish its consequences, allowing individuals to live longer, healthier and independently in old age.

In this paper, we developed risk prediction models based on comprehensive longitudinal data. Methods for prediction based on longitudinal data have been investigated for time-to-event analysis (Maziarz et al., 2017; Sweeting & Thompson, 2011, 2012) and machine learning methods have been applied in similar settings as ours (Choi et al., 2017; Zhao et al., 2019). However, the situation where machine learning has to incorporate right censored outcome data and competing risks has not been explored much. As done by Sweeting et al. (2017), we compared two alternative forms of learning from the longitudinal trajectories. In the first, the predictor variables were summaries of the history of a marker derived by a subject matter expert. The expert-derived variables could then be used with standard regression models as well as with machine learning techniques. In the second form, a complex machine learning algorithm was trained directly with the observed longitudinal trajectories.

Accounting for the right censored observations of our outcome, we implemented the expert-derived predictor variables in standard regression models for competing risks: cause-specific Cox regression (Benichou & Gail, 1990) and Fine–Gray regression (Fine & Gray, 1999). We also adapted artificial neural networks to the competing risk outcome with baseline predictor variables (Lee et al., 2018) and extended recurrent neural networks (Hochreiter & Schmidhuber, 1997) in order to data adaptively derive predictor variables from person specific longitudinal trajectories. Next, we applied the modelling algorithms to register data from the Municipality of Copenhagen and obtained rival prediction models based on the regression and machine learning approaches. We then compared the predictive performance using a temporal validation study. We also investigated to what extent the models built in Copenhagen can be exported to other municipalities in Denmark (spatial validation). Finally, we analysed the relative utility of single data sources by repeating the validation when specific registers were left out of the model.

More generally, in this paper we discuss ways to validate and compare traditional regression models with machine learning models for time to event outcomes with competing risks. We illustrate the general steps of design, data analysis and modelling throughout by predicting the need of home care services as a motivating example. We start by discussing the time dynamics of the design and data collection where the object of the analysis is defined relative to a time zero, an event of interest and

a time horizon. This is followed by a discussion on the challenges of deriving predictor variables from register data where we distinguish the classical modelling culture, where predictor variables are derived based on expert knowledge, from the machine learning modelling culture, where predictor variables are learnt from the data (Breiman, 2007).

## 2 | NATION-WIDE REGISTERS

There are advantages and disadvantages in using nation-wide registers as a data source for the purpose of developing a prediction model. It is an advantage that they do not exclude individuals in a systematic way. It is also an advantage that the main reason for early end of follow-up, apart from administrative censoring, is moving out of the country. Hence, prediction models built on nation-wide registers are potentially applicable for everyone in the country. However, it is a disadvantage that these electronic records do not necessarily contain specific information for the current purpose. For example, information on physical functioning and lifestyle would be useful for predicting the need of home care. Unfortunately, such data are not available through the registers and can only be roughly approximated based on electronic medical and consumption records. Nonetheless, it is possible that an algorithm can extract specific information by searching through a person's records across all available registers.

The start of a register limits how long back in time we have personal information on a given individual. With increasing age of the register, this limitation gets lessened, but generally the records can only be complete for subjects born after the start of the register. The end of the register also has consequences. For example, to predict the outcome within a 5-year horizon, subjects had to be eligible and registered at least 5 years ago. More generally, the longer the prediction horizon, the older the data. Here we restricted to a 17-year history and also compared to a 10-year history using registers with information on demographics, income, medications, hospitalisation, care services, death and immigration/emigration, see Figure S1 for a timeline overview of these registers.

Statistics Denmark is the central authority responsible for maintaining hundreds of high quality Danish registers and producing fine-grained statistics on changes in many aspects of life, for example, social, economic, biomedical conditions and geographical location over time (Frank, 2000). In general there are no missing values in the registers that record services use or consumption. Although these data are largely archived in cross-sectional registers, in 1968 the Danish Civil Registration System established a unique identifier for everyone in the population, making it possible to systematically trace individuals over time and space (Pedersen, 2011). The data used for this project are housed in a protected environment in concordance with national and international guidelines on the use of data for analytical purposes. Identifiable and re-identifiable data can per legal policy not be exported from Statistics Denmark (2017). Here, we used the following registers:

*The Danish Civil Registration System.* All persons alive and living in Denmark are registered by the Danish Civil Registration System, which has existed since 1968 (Schmidt et al., 2014). Here we used information on a person's age, sex, country of origin, immigration and civil status, family and household types and municipality of residence.

*The Population Education Register.* This register contains information on an individual's highest completed education ever since 1981, including main education group, type of education and exact title, which corresponds to The International Standard Classification of Education codes. Aalen & Johansen, 1978, there is information for 96.4% of the Danish population aged 15–69 (Jensen & Rasmussen, 2011). This means there are missing information on education for the oldest old, which we categorised as 'unknown'.

*The Income Statistics Register.* This register provides information for individuals and households regarding the composition of the population's income and allowance, available and updated annually since 1970 (Baadsgaard & Quitzau, 2011). Here we used information on total salary including self-employed and honorary fees, income from wealth such as stocks and investments, social security allowance from the state such as cash benefits and parental leave, national and early retirement pension, private pension and disposable income. Data are also adjusted according to family situation, that is, number of people registered at the same address.

*The National Prescription Register.* Since 1995, this register has recorded detailed information on prescriptions redeemed by Danish residents at community pharmacies, including drugs prescribed to nursing home residents (Pottegård et al., 2016). Here we used information on the purchase date, the number of packages purchased, the Anatomical Therapeutic Chemical (ATC) classification system codes and the volume of packages. For our analysis, we aggregated the ATC codes to third-level classifications (pharmacological subgroups), for example, from A10BA02 (metformin) to A10B (blood glucose lowering drugs).

*The National Patient Register.* As one of the world's oldest nationwide hospital registers, it provides detailed administrative and clinical data on all patient contact in Danish non-psychiatric hospitals since 1977 and psychiatric specialty clinics or hospitals since 1995 (Schmidt et al., 2015). Data includes admission date, discharge date, type of patient and primary diagnoses as World Health Organization International Classification of Diseases version 8 (WHO ICD-8) codes (until 1995) and WHO ICD-10 codes (since 1995). In our analysis we aggregated the ICD-10 codes into chapters, for example, from G30.0 (Alzheimer's disease with early onset) to Chapter 6 (Diseases of the nervous systems).

*The Older Persons Register on Home Care Visits.* This register provides individual-level information on permanent help in the form of care in residents' own homes. Data includes dates of home care visits and the type and duration of home care services provided, recorded ever since 2008. This includes both personal and practical care. Personal care is defined as assistance given to persons who, due to reduced physical or mental disabilities or special social problems, are not able to perform tasks such as assistance with bathing, toilet visits, coming out of the bed, dressing, medication handling and so forth. Practical care is defined as instrumental assistance such as cleaning, food service, shopping, laundry and so forth. We used the dates and total minutes of combined personal and practical home care per month as one of our outcomes.

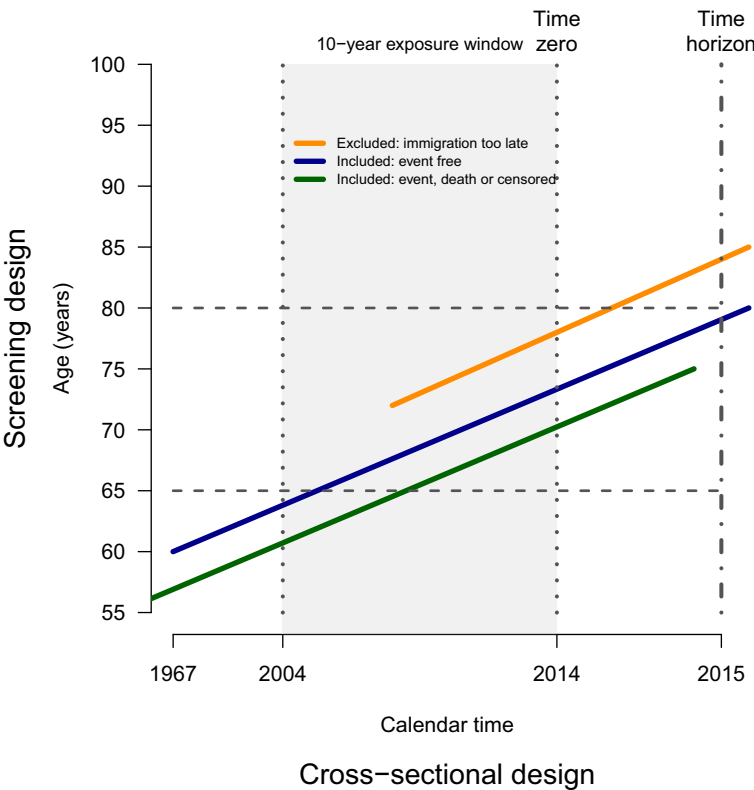
*The Older Persons Register on Nursing Home.* This register records detailed individual data on nursing home admittance ever since 2008. Here, we used dates of the first admittance into nursing home as one of our outcomes.

*The Death Register.* Since 1970, this register has been fully digitalised and includes all deaths of Danish residents dying only in Denmark. Ever since 1983, the registers also include deaths among Greenlanders and Faroese living in Denmark and dying in Denmark, Greenland or the Faroe Islands (Helweg-Larsen, 2011).

### 3 | ANALYSIS DESIGN AND DATA STRUCTURE

#### 3.1 | Analysis design

For our application, we considered two alternative analysis designs: the *screening design* and the *cross-sectional design* (see Figure 1). In the *screening design*, all persons receive a predicted risk when they reach a certain age, say 70 years. The calendar dates at which the model is applied depends



**FIGURE 1** Illustration of the design of the prediction model data analysis. The cross-sectional design predicts the risk of all persons in a given age range at a single calendar date (time zero). The screening design invites all persons who reach a single age value in a given calendar year to get a predicted risk [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

on the birth year of the persons. The predicted risk can then in principle be updated for individual persons when they get older. In the *cross-sectional design*, all persons of the population in a given age range, say 65–80 years, receive a predicted risk at a single calendar date, that we call time zero. The prediction can then be repeated frequently, say once a year. Since our aim is to assist the Danish municipalities in optimising preventive strategies and not to design a screening tool, we only pursued and reported from the *cross-sectional design*.

We used longitudinal register data from an exposure window before time zero to predict event probabilities between time zero and the time horizon (see Figure 1). This is different from time-varying covariates observed during the outcome time, as typically analysed in inferential survival analysis (Kalbfleisch & Prentice, 2002). However, our analysis can be seen as a landmark analysis with a single time point (Blanche et al., 2015; Rizopoulos, 2012). Repeated predictions, as explained above, would add new time points to the landmark analysis to update predictions in a dynamic way.

**3.2 | Outcome event, competing risks and prediction horizon**

The target parameter is the probability that the event of interest occurs within a given time horizon since time zero in any subject of the target population. We call this probability the  $\tau$ -year

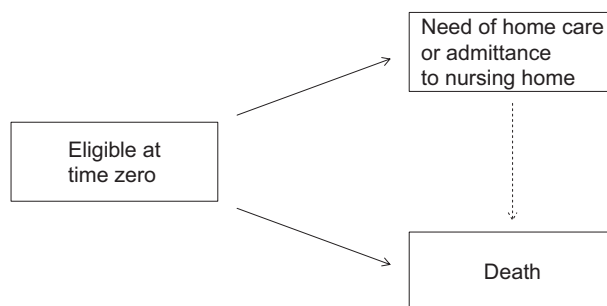
risk of the event for a suitably selected time horizon  $\tau$ . The time horizon is seen from the standpoint of time zero, that is, where the time scale is time since time zero. Thus, we need to define the event of interest and one or several time horizons. In our application, the event of interest is the combined endpoint *need of home care* or *admittance to a nursing home*. More formally, let  $N_1(t)$  be a counting process which at any time  $t$  counts the current need of home care in minutes per month. We assume that  $N_1(t)$  is a monotone increasing process and define a binary outcome variable defined when the home care process exceeds the threshold  $\xi$  at any time before the prediction time horizon  $\tau$ :

$$Y_1(\tau, \xi) = \begin{cases} 0 & N_1(\tau) < \xi \\ 1 & N_1(\tau) \geq \xi. \end{cases}$$

In what follows we choose  $\tau=1$  year after the date of time zero as our prediction time horizon and use  $\xi=60$  min per month to define the event: *need of home care*. Letting  $N_2(t)$  count admittance to a nursing home (0=not in nursing home, 1=in nursing home), we define the binary outcome variable for the combined endpoint as

$$Y(\tau) = 1 \{ (Y_1(\tau, \xi) + N_2(\tau)) > 0 \}.$$

Finally, we need a third counting process  $N_3(t)$  which takes the value zero while the person is alive and the value one when the person died. The processes  $N_1$  and  $N_2$  do not change after death and  $N_3$  is also used to define the target population. The target population contains only subjects who are theoretically at risk of the event. That is, subjects for whom the event occurred before time zero are excluded. See Figure 2 for a visualisation of this multi-state model. At time zero, all subjects from the target population are in the eligible state. At the time horizon, a person can be in the eligible state (no event), in the *Need of home care or admittance to nursing home* state (event of interest) or in the *Death* state (competing event). We do not consider transitions from *Need of home care or admittance to nursing home* to *Death*. Other competing risks could be defined if this improves the interpretation of the predicted risks, but it should be noted that this would usually alter the definition of the event of interest. For example, we could define our event of interest as *need of home care* and consider *admittance to a nursing home* as a competing risk. Then, the interpretation of the predicted risk would be the probability of *need of home care before admittance to a nursing home*.



**FIGURE 2** Multi-state model. At time zero, all persons are in the eligible state. At the time horizon, a person can be in either of the three states

### 3.3 | Predictor variables

To describe the structure of the predictor variables in our case study, we let  $R$  be the number of data sources from which predictor variables are extracted. A longitudinal or time varying variable is a variable which can change over time. We denote by  $X_j^r(s)$  the  $j$ th variable extracted from the  $r$ th register evaluated at time  $s \leq 0$  where 0 is time zero defined above. Longitudinal variables are evaluated at the cross-section with time zero and this current status value,  $X_j^r(0)$ , can be used for prediction. In addition to the current status value at time zero, there may be predictive value in the changes of the process  $X_j^r(s)$  over time. Thus, for longitudinal variables we fix a length of the exposure time window  $w$  and consider as predictors the trajectories in the  $w$ -year history of each person:

$$\mathbf{X}_{j,w}^r = \{X_j^r(s) : w \leq s \leq 0\}.$$

Note that, in principle, one can use different exposure time windows for different variables and that the start of the register  $r$  is a natural upper limit for the value  $w$  (see Section 3.4). We use the notation  $\mathbf{X} = \{\mathbf{X}_{j,w}^r : j \leq J_r, r \leq R\}$  to describe all observed trajectories across all registers in the considered exposure time window.

A derived predictor variable  $Z_k$  is any variable which arises from applying a function  $f_k$  to the ( $w$ -year) history of the other variables (across the registers), that is,

$$Z_k = f_k(\mathbf{X}_{j,w}^r : r = 1, \dots, R; j = 1, \dots, J_r).$$

We consider  $k=1, \dots, K$  derived predictor variables, where  $K$  is an integer which may in fact be larger than the total number of observed variables,  $\sum_{r=1}^R J_r$ . A simple example of a derived variable is when rare categories of a categorical variables are merged somehow. For example, we can collapse the country of origin into a new derived variable which has only two classes:  $\{\text{Denmark}\}$  and  $\{\text{Rest of the world}\}$ . A more complex example would be the number of hospital admissions in the  $w$ -year history which triggered at least one new drug purchase within 8 weeks after the discharge from hospital. Derived variables can either be specified by a priori expert knowledge (classical statistical approach) or learnt from the data in a supervised fashion (machine learning approach).

### 3.4 | Censored observation

In our case study, we have to deal with different censoring patterns before and after time zero. To describe the censoring patterns before time zero, note that the register-specific processes are left-censored at the start of the register. This means for example that we can see that a patient purchases anti-diabetic medicine but the onset of diabetes, that is, the date of the first purchase is unknown (left-censored). We deal with this by limiting the length of the exposure time window  $w$  to be short enough such that the registered history of all contributing persons is equally long. Furthermore, persons are excluded from our analysis (and from being eligible for using our prediction model) if they immigrated into Denmark so late that they do not have the full exposure time window recorded in the Danish registers (see Figure 1). Persons are also excluded if they emigrated out of Denmark at any time point during the exposure time window, even if the immigrated into Denmark again before time zero. This is in order to be able to calculate the full  $w$ -year histories  $\mathbf{X}_{j,w}^r$ . After time zero, we right censor the outcome processes for persons when they emigrate out of Denmark before the prediction time horizon.



Let  $C$  be the right censoring time, and  $T$  the time to need of home care, nursing home or death, whatever comes first. The observed event time is  $\tilde{T} = \min(T, C)$  and the event variable and censoring indicator are  $\Delta = D I\{C \geq \min(T)\}$  where  $D=1$  if either need of home care or nursing home occurs and  $D=2$  if death occurs without need of home care or nursing home. The observation is right censored when  $\Delta=0$  and in this case neither  $T$  nor  $D$  are observed. The observed data set  $D_n = (\mathbf{X}_i, \mathbf{Z}_i, \tilde{T}_i, \Delta_i)_{i=1}^n$  consists of data from  $n$  individual persons who are eligible according to the exposure time window  $w$  and alive (without event) at time zero.

## 4 | MODELLING

In the competing risk setting (Figure 2), the standard survival analysis provides several alternative regression modelling strategies (see e.g. Gerds et al., 2012). Here, we focus on the most popular ones, that is, the cause-specific Cox regression and the Fine–Gray regression. Furthermore, we apply artificial neural networks for competing risks (Lee et al., 2018) and combine these networks with recurrent neural networks (Hochreiter & Schmidhuber, 1997) to learn from longitudinal data. In line with the classical modelling culture, we fit the two regression methods to derived predictor variables  $\mathbf{Z}$  and, in line with the machine learning modelling culture, the two neural network methods to untransformed predictor variables  $\mathbf{X}$ . In the following, we denote by  $M$  a prediction model, and by  $\hat{M}(\mathbf{Z}_{new}, \tau)$  and  $\hat{M}(\mathbf{X}_{new}, \tau)$  the predicted  $\tau$ -year risks for predictor variables of a new individual  $\mathbf{Z}_{new}$  and  $\mathbf{X}_{new}$ , respectively.

### 4.1 | Cause-specific Cox regression

Our first candidate prediction model combines several Cox regression models into a predicted risk of the outcome (Benichou & Gail, 1990; Ozenne et al., 2017). Specifically, we use a multiple Cox regression for the hazard rate of *need of home care or nursing home*:

$$\lambda_1(t|\mathbf{Z}) = \lambda_{01}(t)e^{\{\beta^T \mathbf{Z}\}}$$

where  $\mathbf{Z} = (Z_1, \dots, Z_K)$  are the derived variables,  $\lambda_{01}$  is an unspecified baseline hazard rate and  $\beta = \beta_1, \dots, \beta_K$  a vector of log-hazard ratios. Similarly, we specify another multiple Cox regression for the hazard rate of death

$$\lambda_2(t|\mathbf{Z}) = \lambda_{02}(t)e^{\{\gamma^T \mathbf{Z}\}}$$

Based on these two models we predict the probability of the outcome event for a new person's-based derived predictor values  $\mathbf{Z}_{new}$  with the formula:

$$\hat{M}^{CSC}(\mathbf{Z}_{new}, \tau) = \int_0^\tau \exp \left\{ - \int_0^t [\lambda_1(u|\mathbf{Z}_{new}) + \lambda_2(u|\mathbf{Z}_{new})] du \right\} \lambda_1(t|\mathbf{Z}_{new}) dt.$$

### 4.2 | Fine–Gray regression

The Fine–Gray regression model (Fine & Gray, 1999) provides an alternative, more direct way of modelling the absolute risk of the outcome:



$$\hat{M}^{FGR}(\mathbf{Z}_{new}, \tau) = \exp \left\{ -a_{01}(\tau) \exp \left[ \alpha^T \mathbf{Z}_{new} \right] \right\}.$$

Here,  $a_{01}$  is an unspecified sub-distribution hazard function and  $\alpha$  a  $k$ -vector of sub-distribution hazard ratios.

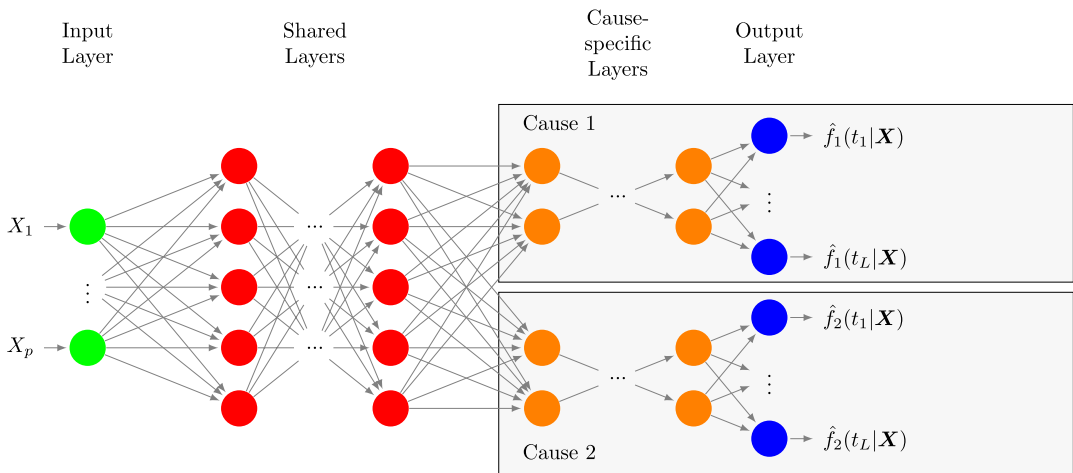
### 4.3 | Neural networks

Several artificial neural network methods for time to event outcomes have been proposed in the last 25 years (e.g. Biganzoli et al., 1998; Faraggi & Simon, 1995; Gensheimer & Narasimhan, 2018; Katzman et al., 2018). However, only few methods are available for competing risks (e.g. Biganzoli et al., 2006; Lee et al., 2018). Here, we adapt the time-discrete setting of Lee et al. (2018) but use a different loss function. Our loss function corresponds to a time-discrete approximation of the direct parameterisation of the cumulative incidence function (Jeong & Fine, 2006). Let  $\{0 = t_1 < t_2 < \dots \leq t_L = \tau\}$  be an equidistant partition of the interval  $[0, \tau]$ . Using a network architecture as illustrated in Figure 3, the output nodes consist of predicted non-conditional event probabilities of events  $q=1,2$  in the time intervals  $[t_{l-1}, t_l)$ ,  $l=1, \dots, L$ , denoted by  $\hat{f}_q(t_l | \mathbf{X})$ . Then, the cumulative incidence function for the event of interest (event 1) can be computed as a simple cumulative sum of the output nodes:

$$\hat{M}^{ANN}(\mathbf{X}_{new}, \tau) = \sum_{l=1}^L \hat{f}_1(t_l | \mathbf{X}_{new}).$$

We implemented the method in the R package `survnnet`, available at <https://github.com/bips-hb/surnnet>.

By construction, these neural networks cannot handle longitudinal input variables such as the  $w$ -year history described above without considering derived predictor variables (Sweeting et al., 2017), as we do with the classical modelling approaches described above. Instead, we use  $X_j^*(0)$ , that is, the



**FIGURE 3** Computational graph of a neural network for competing risks with two causes. The network inputs  $X_1, \dots, X_p$  are numeric vectors. The output is the probabilities of event  $q$  in the time partition  $[t_{l-1}, t_l)$ ,  $l=1, \dots, L$ . [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

longitudinal variables evaluated at time zero and describe recurrent neural networks, a method from the machine learning modelling culture which is able to data adaptively learn derived variables from the  $w$ -year history for prediction, in the next section.

#### 4.4 | Recurrent neural networks

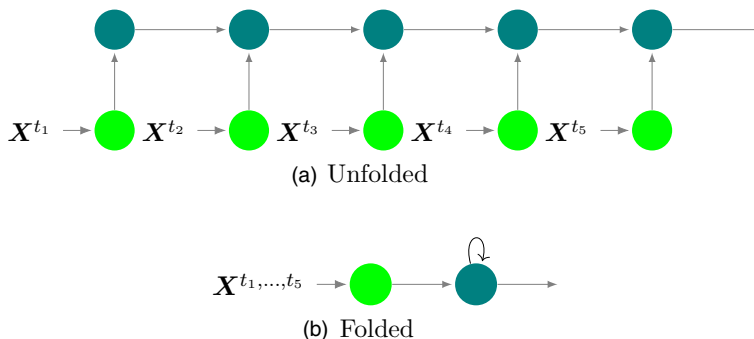
Recurrent neural networks (RNNs) are neural networks that process sequential or longitudinal data. Typically, the covariate data are represented by a three-dimensional array. In our case, we observe several variables for several persons at several time points. Figure 4a shows how this can be incorporated into the input layer of a neural network. For every time interval  $t_i$ , an input unit is created and connected with the previous time interval. Figure 4a shows the so-called unfolded representation of the unit. To simplify computational graphs, these units are often drawn in the folded representation, as shown in Figure 4b.

In principle, one could use standard neural network units with any activation function as units in the RNN. However, this would lead to overrepresentation of the last elements of the sequence because these are more closely connected to the output layer of the network. This phenomenon is called the vanishing gradient problem (Hochreiter et al., 2001) or sometimes short-time memory because the network tends to ‘forget’ the first time intervals. As a consequence, several specific recurrent units have been proposed. The most popular ones are the long short-term memory (LSTM, Hochreiter & Schmidhuber, 1997) and gated recurrent units (GRU, Cho et al., 2014).

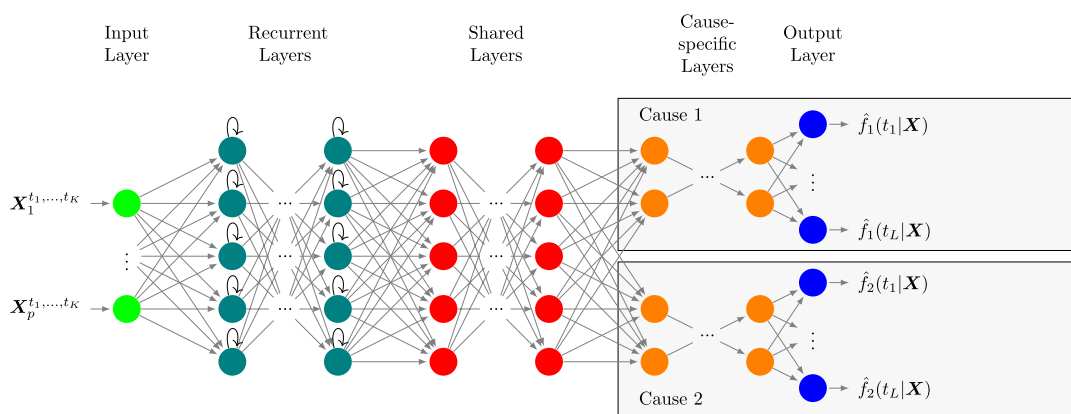
Here, we combined Lee et al. (2018)’s methods for competing risks with recurrent layers and created RNNs for competing risks. We used a combination of recurrent layers, shared layers and cause-specific layers and predict non-conditional event probabilities for each cause and time interval, as described in Section 4.3. See Figure 5 for an overview of the network architecture. The prediction was obtained in the same way as for the non-recurrent neural network:

$$\hat{M}^{RNN}(\mathbf{X}_{new}, \tau) = \sum_{l=1}^L \hat{f}_l(t_l | \mathbf{X}_{new}).$$

We also implemented the RNNs in the R package `survnet`, available at <https://github.com/bips-hb/survnet>.



**FIGURE 4** Computational graph of the unfolded and folded representation of a recurrent unit [Colour figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com)]



**FIGURE 5** Computational graph of a recurrent network for competing risks with two causes. The network inputs  $X_1^{t_1, \dots, t_K}, \dots, X_p^{t_1, \dots, t_K}$  are matrices with observations for time points  $t_1, \dots, t_K$ . The output is the probability of event  $q$  in the time partition  $[t_{l-1}, t_l]$ ,  $l=1, \dots, L$  [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

## 4.5 | Performance metrics and validation

To evaluate the prediction performance, we considered the following time-dependent metrics. The Brier score of a model  $M$  at prediction time horizon is the expected value of the squared difference between the binary outcome of a new person and the predicted risk by the model:

$$\text{Brier}(\tau, M) = E \{ Y_{new}(\tau) - \hat{M}(\mathbf{Z}_{new}, \tau) \}^2.$$

Note that the neural network methods do not use the derived predictor variables and thus  $\mathbf{Z}_{new}$  is replaced by  $\mathbf{X}_{new}$ . We also considered the null model  $M^0$  which ignores the covariates and uses the Aalen–Johansen estimator (Aalen–Johansen, 1978) to estimate the outcome risk. Based on the Brier score for the null model,  $\text{Brier}(\tau, M^0) = E \{ Y_{new}(\tau) - \hat{M}^0(\tau) \}^2$ , we construct the index of prediction accuracy (IPA) for model  $M$  (Kattan & Gerds, 2018):

$$\text{IPA}(\tau, M) = 1 - \frac{\text{Brier}(\tau, M)}{\text{Brier}(\tau, M^0)}.$$

The IPA measures both discrimination and calibration of the model. To look further into the model performance, we also calculated calibration plots (Gerds et al., 2014) and the time-dependent area under the curve (AUC) for the competing risk setting (Blanche et al., 2013). The latter is defined as the probability that a randomly selected person with the event at time  $\tau$  receives a higher predicted risk than a randomly selected person who either died without the event before time  $\tau$  or is event-free at time  $\tau$ :

$$\text{AUC}(\tau, M) = P(\hat{M}(\mathbf{Z}_{new_1}, \tau) > \hat{M}(\mathbf{Z}_{new_2}, \tau) | Y_{new_1}(\tau) = 1, Y_{new_1}(\tau) = 0).$$

As above  $\mathbf{Z}_{new}$  is replaced by  $\mathbf{X}_{new}$  for the neural network methods. To deal with the right censored outcome data, we used inverse probability of censoring weighted estimates of the Brier score and the AUC, described elsewhere (Blanche et al., 2013; Gerds & Schumacher, 2006).

In order to predict how well the models will predict the  $\tau$ -year future of a future person we estimated the prediction performance metrics to assess the models and differences between the models.

The general idea was to first train all candidate models in a learning set of data and to use a validation set of data to estimate the prediction performance metrics. Here, we used temporal validation and spatial validation. With temporal validation, we applied our modelling strategies to the 2014 data of the Municipality of Copenhagen (time zero: 1 January 2014) and validated the selected models in the 2015 data of the Municipality of Copenhagen (time zero: 1 January 2015). With spatial validation, we applied our modelling strategies to the data of the Municipality of Copenhagen and validated the selected models in the data of the Municipality of Aarhus as well as the rest of Denmark. Model development and validation were conducted in accordance with the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) statement (Collins et al., 2015).

## 5 | RESULTS

Here, we present the results of our application to predict the need of future home care services. First, we show descriptive statistics on the population and the event of interest. Second, we compared our candidate models, as described in Section 4, in temporal and spatial validation (see Section 4.5). Third, we analysed the relative utility of single data sources, that is, the registers included in the prediction models.

We compared the cause-specific Cox model (CSC), Fine–Gray regression (FGR), non-recurrent artificial neural network (ANN) and recurrent neural network (RNN). For the RNN, a model using the whole 17 years of personal history (RNN-w17) and a model using the last 10 years history (RNN-w10) were compared. For CSC and FGR, we used the following (derived) predictor variables: age, sex, house type, disposable income quantiles, number of different hospital diagnoses in the year before time zero and number of different prescribed medications in the year before time zero. In the neural networks, we included all available variables without any data preprocessing or feature engineering (see Table S2 for a list of all variables). The hyperparameters of the neural networks were tuned on the Copenhagen (2014) data. We did not look at the validation data before fixing the hyperparameters. The used hyperparameters are shown in Table S1. We also included a benchmark model, which is a CSC with only age and sex as covariates.

### 5.1 | Descriptive statistics

Table 1 shows the population characteristics for Copenhagen in 2014, Aarhus in 2015 and the whole of Denmark in 2014. Figure 6 shows the percentage of persons in the population who used home care or were admitted to a nursing home in the year 2014 according to municipalities in Denmark. Figures S2 and S3 show the percentage of persons in the population who used home care or were admitted to a nursing home in the year 2014 according to age and sex in all of Denmark and the cumulative incidence for the Municipality of Copenhagen between January 2014 and December 2017, respectively. The event rates increased with age and, in old age, were higher for women than for men, whereas mortality was higher for men than for women. Figure 6, Figures S2 and S3 were created using Aalen–Johansen estimates.

### 5.2 | Temporal validation

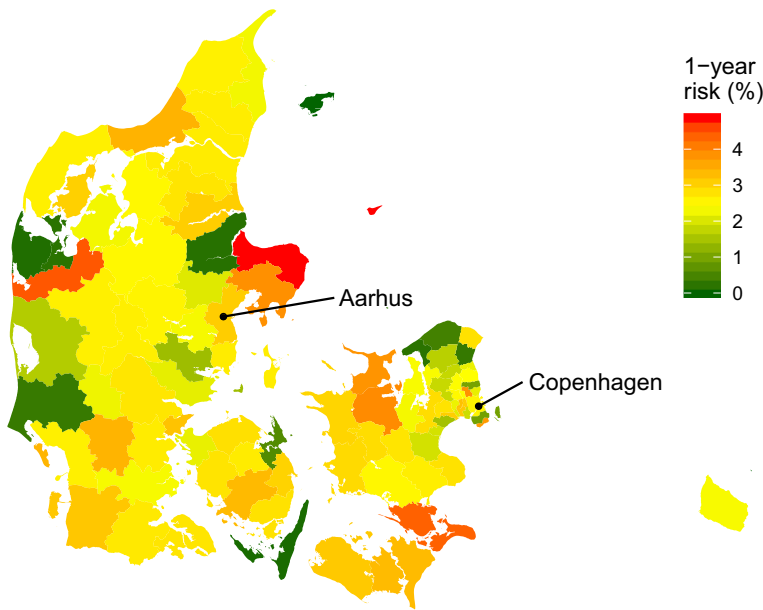
To evaluate the temporal transportability of fitted models, we built models on the Copenhagen data where time zero is 1 January 2014. With these models, we calculated performance metrics using the data of the Copenhagen population for those who were eligible 1 January 2015 to predict need of home care in the following year, that is, until 31 December 2015. Figure S4 shows the predicted 1-year risks,

TABLE 1 Population characteristics

| Variable                      | Level             | Copenhagen (2014) | Aarhus (2015)  | Denmark (2014)  |
|-------------------------------|-------------------|-------------------|----------------|-----------------|
|                               |                   | n=173,280         | n=57,249       | n=1135,865      |
| Sex                           | Female            | 90,812 (52.4%)    | 29,813 (52.1%) | 581,398 (51.2%) |
|                               | Male              | 82,468 (47.6%)    | 27,436 (47.9%) | 554,467 (48.8%) |
| Age at entry                  | 55–65             | 56,642 (32.7%)    | 14,802 (25.9%) | 313,300 (27.6%) |
|                               | 65–70             | 53,128 (30.7%)    | 17,287 (30.2%) | 329,775 (29.0%) |
|                               | 70–75             | 32,406 (18.7%)    | 12,128 (21.2%) | 228,431 (20.1%) |
|                               | 75–80             | 17,084 (9.9%)     | 7,424 (13.0%)  | 145,860 (12.8%) |
|                               | 80–90             | 12,564 (7.3%)     | 5209 (9.1%)    | 109,518 (9.6%)  |
|                               | >90               | 1456 (0.8%)       | 399 (0.7%)     | 8981 (0.8%)     |
| Marital status                | Single man        | 28,154 (16.2%)    | 6036 (10.5%)   | 113,879 (10.0%) |
|                               | Single woman      | 45,874 (26.5%)    | 11,558 (20.2%) | 204,519 (18.0%) |
|                               | Married couple    | 70,884 (40.9%)    | 32,713 (57.1%) | 687,964 (60.6%) |
|                               | Incidental couple | 9814 (5.7%)       | 3029 (5.3%)    | 57,340 (5.0%)   |
|                               | Other             | 18,554 (10.7%)    | 3913 (6.8%)    | 72,163 (6.4%)   |
| Number of drugs purchased     | 0                 | 23,400 (13.5%)    | 6616 (11.6%)   | 134,405 (11.8%) |
|                               | 1                 | 20,578 (11.9%)    | 6507 (11.4%)   | 129,636 (11.4%) |
|                               | 2                 | 20,794 (12.0%)    | 6717 (11.7%)   | 135,796 (12.0%) |
|                               | 3–5               | 55,322 (31.9%)    | 18,761 (32.8%) | 370,917 (32.7%) |
|                               | 6–10              | 44,320 (25.6%)    | 15,580 (27.2%) | 303,368 (26.7%) |
|                               | 11–20             | 8820 (5.1%)       | 3054 (5.3%)    | 61,472 (5.4%)   |
|                               | >20               | 46 (0.0%)         | 14 (0.0%)      | 271 (0.0%)      |
| Number of hospital admissions | 0                 | 84,634 (48.8%)    | 27,421 (47.9%) | 504,441 (44.4%) |
|                               | 1                 | 48,654 (28.1%)    | 15,888 (27.8%) | 337,369 (29.7%) |
|                               | 2                 | 23,524 (13.6%)    | 8301 (14.5%)   | 175,007 (15.4%) |
|                               | 3                 | 10,008 (5.8%)     | 3569 (6.2%)    | 74,395 (6.5%)   |
|                               | 4–5               | 5536 (3.2%)       | 1824 (3.2%)    | 39,214 (3.5%)   |
|                               | >5                | 924 (0.5%)        | 246 (0.4%)     | 5439 (0.5%)     |

comparing CSC, FGR, ANN, RNN-w17 and RNN-w10. CSC and FGR predicted similar risks (A), however, the CSC risks were systematically higher as slightly more dots were below the diagonal. Comparing the CSC risks with the ANN risks (B) shows that individuals can have very different risk predictions by the two models. The same is true for the comparison between ANN and RNN-w17 (C). However, the deviations of RNN-w10 risk to RNN-w17 risk did not have a large magnitude for the majority of the population (D). Our analyses also show that the performance of FGR was very similar to the performance of CSC and that RNN-w17 had slightly worse performance than RNN-w10, see Figure S5. Based on these observations, we did not further consider FGR and RNN-w17 and restricted the remaining analyses to the models Benchmark, CSC, ANN and RNN-w10.

Figure 7 (left) shows the AUC and IPA (see Section 4.5) over 12 months, beginning 1 January 2015. We see that CSC outperformed the benchmark model as much as it was outperformed by ANN which in turn was slightly outperformed by the RNN using a 10-year time history window. All models



**FIGURE 6** Percentage of persons in the population who used home care or were admitted to a nursing home within 1 year since January 2014 according to municipalities in Denmark [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

show a good calibration for risks below about 25%, but on average overestimate risks above 25%. Note that only few observations with risk above 50% existed and most observations were below 25%.

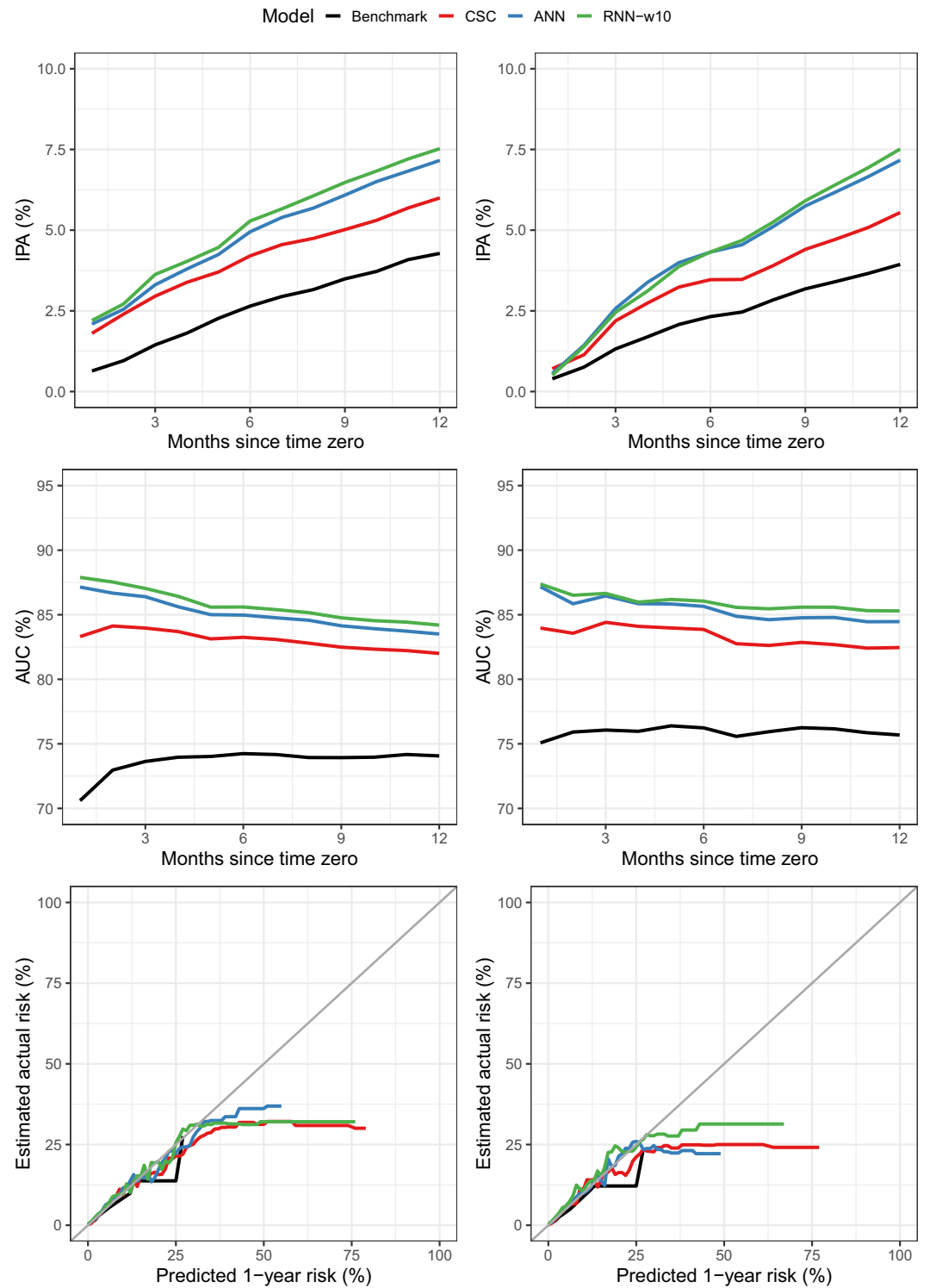
### 5.3 | Spatial validation

Next, we evaluated the spatial transportability of prediction models by building models in Copenhagen municipality and evaluating them in Aarhus municipality. We used the models built in the previous section with time zero at 1 January 2014 and calculated performance metrics on the data of the Aarhus population eligible 1 January 2015. See Figure 7 (right) for the results. The results are mostly similar to the temporal validation described in Section 5.2.

In particular, calibration is important in the attempt to transport a model from one region to another. Even though the results for Aarhus were very promising, a map of the whole country in terms of IPA (Figure 8) shows that in some regions the IPA was negative. This means that a very simple model, which predicted the same risk to every subject, outperformed the recurrent network model. On further reflection, a similar pattern is present with the CSC (see Figure S6) and note that the municipalities with negative IPA matched the ones with very low observed risk (compare Figure 6).

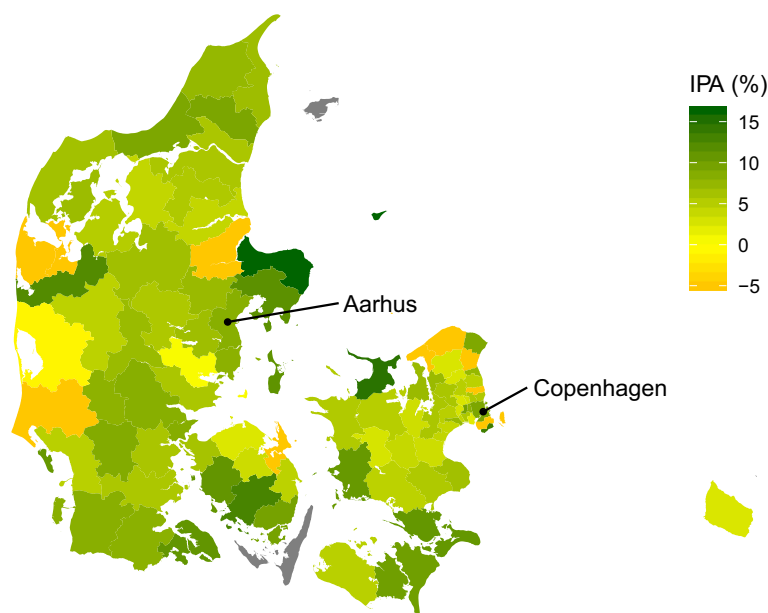
### 5.4 | Relative utility of single data sources

To evaluate the importance of the registers described in Section 2, we repeated the temporal validation as described in Section 5.2 but left out single data sources. Figure 9 shows the difference in AUC, when a given register was left out. We see that the CSC strongly relied on demographics, that is, age and house type, whereas the RNN shows higher importance of medications and hospital admissions.

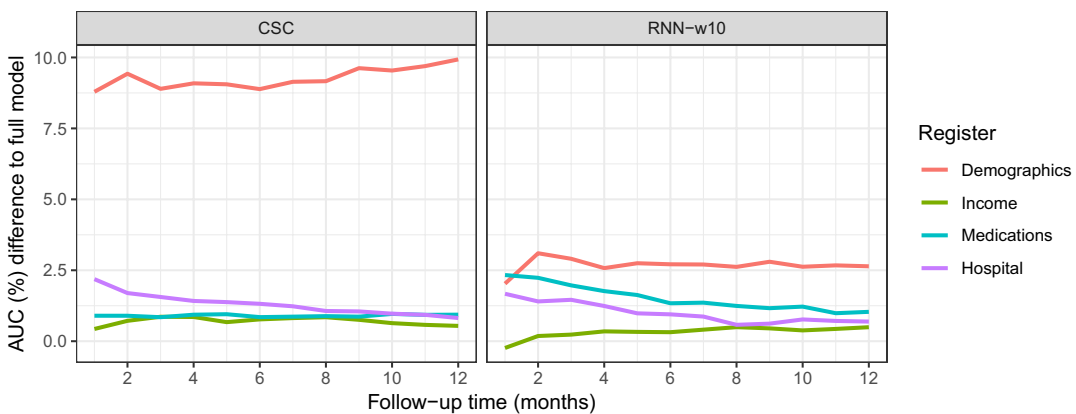


**FIGURE 7** Results of the temporal validation (left) and spatial validation (right), showing IPA (top), AUC (centre) and calibration (bottom) of the benchmark model, cause-specific Cox model (CSC), artificial neural network (ANN) and recurrent neural network using 10 years of personal history (RNN-w10) [Colour figure can be viewed at [wileyonlinelibrary.com](#)]





**FIGURE 8** Validation of RNN (10-year window, trained in Municipality of Copenhagen) in the rest of Denmark [Colour figure can be viewed at [wileyonlinelibrary.com](#)]



**FIGURE 9** Results of the temporal validation when single data sources were left out. For the CSC, *Demographics* consists of the variables age and house type, *Income* of disposable income quantiles, *Medications* of the number of prescribed medications and *Hospital* of the number of hospital diagnoses. For the RNN, these correspond to the whole registers [Colour figure can be viewed at [wileyonlinelibrary.com](#)]

The importance of medications and hospital admissions decreased over time, while the importance of income increased.

## 6 | DISCUSSION

Using Danish register data, we improved neural networks for competing risks (Lee et al., 2018) by combining them with RNNs to predict the risk of needing long time care services (home care or

nursing home) in the future with death as competing risk and thus identify older individuals who are at high risk of declining health. We then assessed its performance by comparing our improved RNN with other established models. Temporal validation indicates the CSC outperformed the benchmark model as much as it was outperformed by the non-recurrent neural network (ANN), which in turn was outperformed by our RNN using a 10-year time history window. While the spatial validation from Copenhagen to Aarhus municipality showed promising results, comparisons across the whole country indicated that in some municipalities with very low observed risk, a very simple model that predicted the same risk to every individual, outperformed our improved RNN. This indicates that these municipalities are structurally different from the other municipalities in Denmark. In terms of relative utility of single data sources, CSC heavily relied on demographics, while the RNN assigned higher importance to medications and hospital admissions. Remarkably, the importance of health information decreased over time, whereas the importance of income increased.

The outcome of interest that we have modelled was either receiving home care for at least 1 hour per month or admittance to a nursing home, whichever comes first, with death as competing risk. From a care provider standpoint, if a person dies earlier than expected, it can be seen as a saving, as a long-term care trajectory is prevented (Taniguchi et al., 2018). Therefore, in terms of prediction, the classification as a negative is correct as there is no use of long-term care. Although correlated, the distinction of these outcomes is of vital interest to correctly predict long-term care services provisions. Moreover, prevention of care services dependency necessitates a different intervention than prevention of mortality (Connolly et al., 2016). When longevity would be the primary interest, a screening model would be more appropriate, whereas mortality would be the primary outcome instead of a competing risk. From a care provider's perspective, there is further complexity as older persons transition between care services, that is, there is a vast heterogeneity in care trajectories with different services and intensity (Colmorten et al., 2004). For example, some may require long term yet simple instrumental help, for example, food delivery, without ever needing the more resource-heavy personal medical care, while some may not need care at all until they are nearing death (Taniguchi et al., 2018). Therefore, the next logical step would be to analyse and predict the various types of care trajectories and what kind of predictors distinguish the one from the other. This is particularly relevant for municipalities since they have to target-specific high risk groups and prioritise which kind of services they should allocate their resources into.

This new way of using data to shape health care strategies introduces new ways of understanding and approaching health and disease (Beam & Kohane, 2018). Furthermore, it also reveals an unforeseen opportunity to compare practice performances, for example, the various preventive programmes the municipalities have to offer. Here, we found considerable differences between the municipalities in the risk of receiving home care or admittance to nursing home. This could be due to differences in health status between the citizens, the social cohesion of people in the community helping each other when in need, how the municipalities organise care or how need of care is reported by the different care providers (Kjaer & Siren, 2019). Whereas most of these differences in the need of care should be picked up by our modelling, the latter results into differences in data quality among the municipalities and may hamper validation and extrapolation of the prediction model. This kind of insight instigates further attention to differences in reporting as these register data were not recorded with the purposes of predicting need of health care services (Schmidt et al., 2019).

It should be emphasised that correctly identifying older persons at risk of long-term care services is just the first step. The obvious next step is to optimise the municipalities' interventions to keep older persons independent and prevent dependency on care services. This is a pertinent goal for most older persons as well as the service provider. The effectiveness of better tailoring the presumed effective interventions however can only be studied using clinical trials (Moons et al., 2012). Note that the

moment these tailored interventions become effective in delaying functional decline and consequently need of health care services, the outcome of the prediction model changes as well, which necessitates continued updating, calibrations and validations of the prediction tool (Lenert et al., 2019).

While these prediction tools hold the promise of optimising preventive interventions in daily practice, ethical concerns with regards to privacy, confidentiality and control have arisen (Price & Cohen, 2019). Since Statistics Denmark operates under the Danish law with the overarching principle to protect the identity of the persons, the use of register data necessitates that privacy and confidentiality are guaranteed (Statistics Denmark, 2017). Effectively, the modelling that we reported here could be performed within this legal and regulatory framework, however the future implementation of this prediction tool requires further ethical considerations (Char et al., 2018). The principal hurdle to overcome is that the municipality as the responsible care provider is not allowed to run an analytic algorithm on an individual's register data to determine risk without explicit consent from the individual itself. Therefore, this risk estimation should be included as an option when offering care services and only used given the individual's consent. With regards to control, a valid prediction tool can only serve as an advisory to help the municipality to target their prevention strategies but not to make decision on prevention or care service delivery (Ngiam & Khor, 2019). Decision making involves respect for citizens' autonomy and consent. Using the prediction tool, the municipality can make a better offering of their preventive interventions and it is up to the individual to decide on whether or not they would accept. Arguably, this is the best way to utilise machine learning methods, at the very beginning as a preliminary assessment, taking into account a plethora of possibilities and providing evidence that would otherwise have not been obtainable given limited time and resources (Schwalbe & Wahl, 2020).

The unique combined strength of the present work lies in developing novel methods capable of handling the complexity of large, pre-existing longitudinal nationwide register data to pre-emptively help older persons maintain their functions thereby alleviating the burden of a societal problem. The novel use of our modelling exercise improved the predictive accuracy and calibration, which provided the municipalities with a tool to optimise their prevention strategies. Although there is traditionally a great emphasis on the predictive values of physical health in use of health care services, the use of care is apparently dependent on socio-demographic characteristics, some of which are readily available in the registers with significant information load. It needs to be explored which of these could ideally be detailed to improve the tool's prediction performance. Although register data are considered as highly valued resource for research, still it is important to understand that its purpose of gathering data is administrative and not for scientific endeavours and or practical applications (Schmidt et al., 2019). Consequently, there may be biases, for example, potential misclassification due to residents in need of help and care was offered by the municipality, but the residents simply rejected the offer (Rostgaard, 2011). Even though such cases are not the norm, nonetheless this means that our outcome could be more appropriately referred as 'use of care' and not 'need of care'. Furthermore, it is a limitation that we did not include primary care data, for example, from general practitioners or physiotherapists as predictors in this study. Currently, these data are only available in a very coarse format as the number of contacts to the physician without a diagnosis code or treatment procedure. As a result, these variables did not pick up any signal in addition to hospital, medication and care data.

In conclusion, taking advantage of the high-dimensionality of Danish population registers, we created further value out of available data by developing a machine learning tool to address a societal problem due to population ageing. We found that our method is capable to accurately predict the use of care services with favourable performance when compared to other established methods, though there were limitations in some regions. Although results so far have been promising, further methodological explorations and ethical considerations leading to clinical experiments are essential to implement the tool.

## ACKNOWLEDGEMENTS

This research was supported by grants from the Novo Nordisk Foundation Challenge Programme: Harnessing the Power of Big Data to Address the Societal Challenge of Aging [NNF17OC0027812].

## REFERENCES

- Aalen, O.O. & Johansen, S. (1978) An empirical transition matrix for non-homogeneous Markov chains based on censored observations. *Scandinavian Journal of Statistics*, 5, 141–150.
- Baadsgaard, M. & Quitzau, J. (2011) Danish registers on personal income and transfer payments. *Scandinavian Journal of Public Health*, 39, 103–105.
- Beam, A.L. & Kohane, I.S. (2018) Big data and machine learning in health care. *Journal of American Medical Association*, 319, 1317–1318.
- Benichou, J. & Gail, M.H. (1990) Estimates of absolute cause-specific risk in cohort studies. *Biometrics*, 46, 813–826.
- Biganzoli, E., Boracchi, P., Mariani, L. & Marubini, E. (1998) Feed forward neural networks for the analysis of censored survival data: A partial logistic regression approach. *Statistics in Medicine*, 17, 1169–1186.
- Biganzoli, E.M., Boracchi, P., Ambrogi, F. & Marubini, E. (2006) Artificial neural network for the joint modelling of discrete cause-specific hazards. *Artificial Intelligence in Medicine*, 37, 119–130.
- Blanche, P., Dartigues, J.-F. & Jacqmin-Gadda, H. (2013) Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Statistics in Medicine*, 32, 5381–5397.
- Blanche, P., Proust-Lima, C., Loubère, L., Berr, C., Dartigues, J.-F. & Jacqmin-Gadda, H. (2015) Quantifying and comparing dynamic predictive accuracy of joint models for longitudinal marker and time-to-event in presence of censoring and competing risks. *Biometrics*, 71, 102–113.
- Borger.dk (2019) Forebyggende hjemmebesøg. Available from: <https://www.borger.dk/aeldre/hjemmehjaelp-og-aeldrepleje/forebyggende-hjemmebesoeg>. Accessed 2019-11-18.
- Breiman, L. (2007) Statistical modeling: The two cultures. *Statistical Science*, 16, 199–215.
- Char, D., Shah, N. & Magnus, D. (2018) Implementing machine learning in health care—addressing ethical challenges. *Massachusetts Medical Society*, 378, 981–983.
- Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. et al. (2014) Learning phrase representations using RNN encoder-decoder for statistical machine translation. *Preprint arXiv:1406.1078*. Available from: <http://arxiv.org/abs/1406.1078>.
- Choi, E., Schuetz, A., Stewart, W.F. & Sun, J. (2017) Using recurrent neural network models for early detection of heart failure onset. *Journal of the American Medical Informatics Association*, 24, 361–370.
- Collins, G.S., Reitsma, J.B., Altman, D.G. & Moons, K.G.M. (2015) Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. *BMJ*, 350, g7594.
- Colmorton, E., Clausen, T. & Bengtsson, S. (2004) Providing integrated health and social care for older persons in Denmark. In: Leichsenring, K. & Alaszewski, A.M. (Eds.) *Providing integrated health and social care for older persons: A European overview of issues at stake*. London: Routledge, pp. 139–179.
- Connolly, D., Garvey, J. & McKee, G. (2016) Factors associated with adl/iadl disability in community dwelling older adults in the Irish longitudinal study on ageing (TILDA). *Disability and Rehabilitation*, 39, 809–816.
- Faraggi, D. & Simon, R. (1995) A neural network model for survival data. *Statistics in Medicine*, 14, 73–82.
- Fine, J. & Gray, R.J. (1999) A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, 94, 496–509.
- Frank, L. (2000) When an entire country is a cohort. *Science*, 287, 2398–2399.
- Gensheimer, M.F. & Narasimhan, B. (2018) A scalable discrete-time survival model for neural networks. *PeerJ*, 7, e6257.
- Gerds, T.A. & Schumacher, M. (2006) Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biometrical Journal*, 48, 1029–1040.
- Gerds, T.A., Scheike, T.H. & Andersen, P.K. (2012) Absolute risk regression for competing risks: Interpretation, link functions, and prediction. *Statistics in Medicine*, 31, 3921–3930.
- Gerds, T.A., Andersen, P.K. & Kattan, M.W. (2014) Calibration plots for risk prediction models in the presence of competing risks. *Statistics in Medicine*, 33, 3191–3203.
- Helweg-Larsen, K. (2011) The Danish register of causes of death. *Scandinavian Journal of Public Health*, 39, 26–29.

- Hochreiter, S. & Schmidhuber, J. (1997) Long short-term memory. *Neural Computation*, 9, 1735–1780.
- Hochreiter, S., Bengio, Y., Frasconi, P. & Schmidhuber, J. (2001) Gradient flow in recurrent nets: The difficulty of learning long-term dependencies. In: Kolen, J.F. & Kremer, S.C. (Eds.) *A field guide to dynamical recurrent neural networks*. New York: Wiley-IEEE Press.
- Jensen, V.M. & Rasmussen, A.W. (2011) Danish education registers. *Scandinavian Journal of Public Health*, 39, 91–94.
- Jeong, J.-H. & Fine, J. (2006) Direct parametric inference for the cumulative incidence function. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 55, 187–200.
- Kalbfleisch, J.D. & Prentice, R.L. (2002) *The statistical analysis of failure time data*, 2 edn. Hoboken, NJ: John Wiley & Sons.
- Kattan, M.W. & Gerds, T.A. (2018) The index of prediction accuracy: An intuitive measure useful for evaluating risk prediction models. *Diagnostic and Prognostic Research*, 2, 7.
- Katzman, J.L., Shaham, U., Cloninger, A., Bates, J., Jiang, T. & Kluger, Y. (2018) DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18, 24.
- Kjaer, A.A. & Siren, A. (2019) Formal and informal care: Trajectories of home care use among Danish older adults. *Ageing & Society*, 40, 1–24.
- Lee, C., Zame, W. & Yoon, J. (2018) Deephit: A deep learning approach to survival analysis with competing risks. In: *AAAI conference on artificial intelligence*. New Orleans, USA.
- Lenert, M.C., Matheny, M.E. & Walsh, C.G. (2019) Prognostic models will be victims of their own success, unless... *Journal of the American Medical Informatics Association*, 26, 1645–1650.
- Maziarz, M., Heagerty, P., Cai, T. & Zheng, Y. (2017) On longitudinal prediction with time-to-event outcome: Comparison of modeling options. *Biometrics*, 73, 83–93.
- Moons, K., Kengne, A., Grobbee, D., Royston, P., Vergouwe, Y., Altman, D.G. et al. (2012) Risk prediction models: II. External validation, model updating, and impact assessment. *BMJ Heart*, 98, 691–698.
- Ngiam, K.Y. & Khor, I.W. (2019) Big data and machine learning algorithms for health-care delivery. *The Lancet Oncology*, 20, e262–e273.
- Ozenne, B., Sørensen, A.L., Scheike, T., Torp-Pedersen, C. & Gerds, T.A. (2017) riskRegression: Predicting the risk of an event using Cox regression models. *R Journal*, 9, 440–460.
- Pedersen, C.B. (2011) The Danish civil registration system. *Scandinavian Journal of Public Health*, 39, 22–25.
- Pottegård, A., Schmidt, S.A.J., Wallach-Kildemoes, H., Sørensen, H.T., Hallas, J. & Schmidt, M. (2016) Data resource profile: The Danish national prescription registry. *International Journal of Epidemiology*, 46, 798–798f.
- Price, W. & Cohen, I. (2019) Privacy in the age of medical big data. *Nature Medicine*, 25, 37–43.
- Rizopoulos, D. (2012) *Joint models for longitudinal and time-to-event data: With applications in R*. Boca Raton: CRC Press.
- Rostgaard, T. (2011) Care as you like it: The construction of a consumer approach in home care in Denmark. *Nordic Journal of Social Research*, 2, 54–69. Available from: <https://doi.org/10.7577/njsr.2042>.
- Schmidt, M., Pedersen, L. & Sørensen, H.T. (2014) The Danish civil registration system as a tool in epidemiology. *European Journal of Epidemiology*, 29, 541–549.
- Schmidt, M., Schmidt, S.A.J., Sandegaard, J.L., Ehrenstein, V., Pedersen, L. & Sørensen, H.T. (2015) The Danish national patient registry: A review of content, data quality, and research potential. *Clinical Epidemiology*, 7, 449–490.
- Schmidt, M., Schmidt, S.A.J., Adelborg, K., Sundbøll, J., Laugesen, K., Ehrenstein, V. et al. (2019) The Danish health care system and epidemiological research: From health care contacts to database records. *Clinical Epidemiology*, 11, 563–591.
- Schwalbe, N. & Wahl, B. (2020) Artificial intelligence and the future of global health. *The Lancet*, 395, 1579–1586.
- Statistics Denmark. (2017) Data confidentiality. Available from: <https://www.dst.dk/en/OmDS/lovgivning>. Accessed 2019-11-18.
- Sweeting, M.J. & Thompson, S.G. (2011) Joint modelling of longitudinal and time-to-event data with application to predicting abdominal aortic aneurysm growth and rupture. *Biometrical Journal*, 53, 750–763.
- Sweeting, M. & Thompson, S. (2012) Making predictions from complex longitudinal data, with application to planning monitoring intervals in a national screening programme. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 175, 569–586.
- Sweeting, M.J., Barrett, J.K., Thompson, S.G. & Wood, A.M. (2017) The use of repeated blood pressure measures for cardiovascular risk prediction: A comparison of statistical models in the ARIC study. *Statistics in Medicine*, 36, 4514–4528.

- Taniguchi, Y., Kitamura, A., Nofuji, Y., Ishizaki, T., Seino, S., Yokoyama, Y. et al. (2018) Association of trajectories of higher-level functional capacity with mortality and medical and long-term care costs among community-dwelling older Japanese. *The Journals of Gerontology: Series A*, 74, 211–218.
- Zhao, J., Feng, Q., Wu, P., Lupu, R.A., Wilke, R.A., Wells, Q.S. et al. (2019) Learning from longitudinal data in electronic health record and genetic data to improve cardiovascular event prediction. *Scientific Reports*, 9, 1–10.

## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

**How to cite this article:** Wright MN, Kusumastuti S, Mortensen LH, Westendorp RG, Gerds TA. Personalised need of care in an ageing society: The making of a prediction tool based on register data. *J R Stat Soc Series A*. 2021;184:1199–1219. <https://doi.org/10.1111/rssa.12644>