Longitudinal Healthcare Analytics for Disease Management: Empirical Demonstration for Low Back Pain

Author(s):
Müller-Peltzer, Michael; Feuerriegel, Stefan; Vach, Werner; Molgaard Nielsen, Anne; Kongsted, Alice; Neumann, Dirk

Publication date:
2020-05

Permanent link:
https://doi.org/https://doi.org/10.3929/ethz-b-000401179

Originally published in:
Decision Support Systems 132, https://doi.org/10.1016/j.dss.2020.113271

Funding acknowledgement:
- Data-driven health management ()

Decision Support Systems 132 (2020) 113271

Contents lists available at ScienceDirect

Decision Support Systems

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

Longitudinal healthcare analytics for disease management: Empirical demonstration for low back pain

Michael Mueller-Peltzer[a,*], Stefan Feuerriegel[b], Anne Molgaard Nielsen[c], Alice Kongsted[c], Werner Vach[d], Dirk Neumann[a]

[a] University of Freiburg, Platz der Alten Synagoge, 79098 Freiburg im Breisgau, Germany
[b] ETH Zurich, Weinbergstr. 56/58, 8092 Zurich, Switzerland
[c] University of Southern Denmark, Campusvej 55, 5230 Odense, Denmark
[d] University Hospital Basel, Spitalstr. 21, 4031 Basel, Switzerland

ARTICLE INFO

Keywords:
Healthcare analytics
Disease management
Longitudinal monitoring
Time series analysis
Cohort data
Low back pain

ABSTRACT

Clinician guidelines recommend health management to tailor the form of care to the expected course of diseases. Hence, in order to decide upon a suitable treatment plan, health professionals benefit from decision support, i.e., predictions about how a disease is to evolve. In clinical practice, such a prediction model requires interpretability. Interpretability, however, is often precluded by complex dynamic models that would be capable of capturing the intrapersonal variability of disease trajectories. Therefore, we develop a cross-sectional ARMA model that allows for inference of the expected course of symptoms. Distinct from traditional time series models, it generalizes to cross-sectional settings and thus patient cohorts (i.e., it is estimated to multiple instead of single disease trajectories). Our model is evaluated according to a longitudinal 52-week study involving 928 patients with low back pain. It achieves a favorable prediction performance while maintaining interpretability. In sum, we provide decision support by informing health professionals about whether symptoms will have the tendency to stabilize or continue to be severe.

1. Introduction

Many diseases afflict a patient's health in the long run. Examples include chronic diseases, such as asthma, arthritis, and epilepsy [13], and other long-lasting conditions, such as headache disorders, depression, and low back pain [39]. The long-lasting nature of these conditions requires health management to closely monitor the progression of the disease and adjust the form of care according to the expected course. This underlies general guidelines for disease management in practice [23].

In order to facilitate effective health management, prior literature has developed various approaches for providing decision support [57]: One major line of research focuses on risk scoring, where the probability of patient-specific health outcomes is estimated. Examples of such outcomes are mortality (e.g., [19]), hospitalization events (e.g., [45]), and hospital readmissions (e.g., [3,7,64]), but not the expected short-term progression of a disease. A different line of research is concerned with the design of treatment plans, namely the timing (e.g., [15]) and dosage of medication (e.g., [8,9,47]). However, no decision support regarding the choice of the underlying treatment plan is provided here either.

Modeling the expected course of a condition is important in disease management for multiple reasons. First, it allows health professionals to predict how the disease is expected to evolve and thus provides decision support regarding the choice of the treatment plan [18,60]. For instance, if a severe disease conditions are likely to persist, then it is necessary to adjust the treatment plan to a more intense regimen at an early point in time. If symptoms are likely to decline and stabilize, then health professionals might be inclined to switch to a treatment plan with lower intensity. This might avoid treatments that are otherwise not necessary. Second, patients benefit from knowing the expected course as they can adapt their daily activities accordingly and self-manage. For instance, it allows them to obtain estimates about the extent to which they can work, travel, or need to pursue specific exercise. Third, considering the expected course of a disease in clinical decision-making helps in personalizing treatment plans to the intrapersonal variability of symptoms from patients [23,30]. This is known to ensure effective care. However, the intrapersonal variability of symptoms is also the reason why health professionals struggle in making accurate inferences regarding the future course of a disease

* Corresponding author.
E-mail address: m.mueller.peltzer@gmail.com (M. Mueller-Peltzer).

https://doi.org/10.1016/j.dss.2020.113271
Received 14 June 2019; Received in revised form 19 February 2020; Accepted 20 February 2020
Available online 10 March 2020

[46] and subsequently could benefit greatly from rigorous models.

Inferring the expected course of disease has been formalized previously by different models. While machine learning models have been used for this purpose (e.g., [26,27,58]), the aforementioned machine learning models act in a black-box manner and thus their interpretability is precluded. However, a key requirement of decision support in clinical settings is interpretability [54]: it ensures accountability for high-stakes decisions and, furthermore, allows researchers to validate the estimated models against state-of-the-art findings from clinical research. Time series models may offer the desired level of interpretability, but those that have been tailored to modeling cohort-wide disease progressions are notably lacking.

Our work provides decision support to health management by modeling expected courses of symptoms. For this purpose, we formed an interdisciplinary team with health researchers and have proposed the use of cross-sectional time series models; specifically, a cross-sectional autoregressive moving-average (ARMA) model. This model predicts the short-term progression and, on top of that, fulfills three requirements from clinical practice: (1) It yields a parsimonious model specification and thus warrants interpretability. (2) Its specification is theory-informed by modeling the intrapersonal variability of symptoms (e.g., pain) via autoregression [46]. (3) It is cross-sectional, so that its parameter estimates generalize to patient cohorts. Thereby, it differs from traditional time series analysis where one model is fitted to one time series. In our model, we fit one model to multiple time series. Later, we also study several extensions (e.g., higher-order lag structures, volatility clusters, and cohort-specific subgroups).

The proposed approach is evaluated based on low back pain, which ranks as the number one cause of high rates of years lived with disability (YLD) rates globally [39]. The course of this disease is particularly characterized by considerable variability, such as short-term pain attacks and subsequent mean-reversals [40,41]. The variable course of pain is illustrated in Fig. 1. For our research, we draw upon longitudinal data from 928 patients with low back pain over a time period of 52 weeks. Our model predicts the next-week pain levels out-of-sample with an RMSE of 1.41. This is then compared with alternative model specifications, which, despite more complex structure, are largely on par in terms of performance. Finally, we discuss the estimation results of our model in order to demonstrate its interpretability.

Our work has multiple implications for disease management. First, our model is aimed to provide decision support for health professionals: they can infer the expected course of a condition and adjust the form of care accordingly. Consequently, practitioners can more precisely explain the most likely course of symptoms to individual patients and thus inform their self-management plans (cf. [25,31]). Second, we extend cross-sectional models, so that they cater for non-contiguous time series. We also propose a novel cross-sectional GARCH (generalized autoregressive conditional heteroskedastic) model. Third, clinical research has noted that cohort-wide disease dynamics are often not appropriately modeled [34,46]. A remedy is provided by our model, as it helps clinical researchers in better understanding the progression of diseases across patient cohorts. Finally, the generic specification of our model ensures broad applicability to other long-lasting diseases.

The rest of this paper is structured as follows. Section 2 summarizes previous efforts to develop decision support for disease management. Section 3 contributes to this by modeling the expected course of diseases in cohort-wide settings. The model is then applied to data from our longitudinal study of 928 patients with low back pain as described in Section 4. The prediction performance is evaluated in Section 5, in which we also report estimation results to demonstrate interpretability. Based on this, Section 6 derives implications for both research and decision-making in clinical practice.

2. Background

By making inferences from patient data, healthcare analytics has the capacity to personalize decision-making in disease management to the specific health profiles of individuals. It thereby provides decision support that promises to be more effective than the usual “one-fits-all” paradigm in care [30]. Prior literature has been concerned with different objectives that can be loosely grouped into (i) risk scoring, (ii) optimizing the design of a given treatment plan, and (iii) predicting disease progression for deciding upon treatment plans. These are detailed as follows.

First, risk scoring predicts the probability of patient-specific health outcomes. Examples of health outcomes include the probability of, for instance, recovery (e.g., [36]), mortality (e.g., [19,65]), onset of a certain disease (e.g., [52]), hospitalization events (e.g., [45]), and hospital readmissions (e.g., [3,7,64]). Methodologically, this is usually formalized in either probabilistic models, such as survival models (e.g., [7]), or machine learning models, such as decision trees (e.g., [52]) or recurrent neural networks (e.g., [4]). On the basis of these models, risk scoring computes the probability that certain, predetermined events in a patient's trajectory take place (e.g., complete recovery or death). This

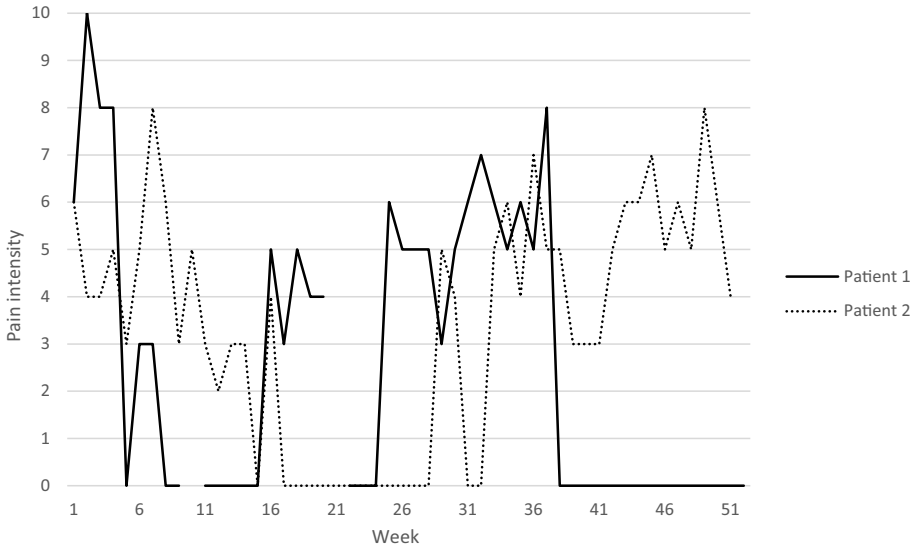

Fig. 1. Examples of the progression of weekly pain from two random patients in our longitudinal study. Pain intensity ranks from 0 “no pain” to 10 “worst imaginable pain”.

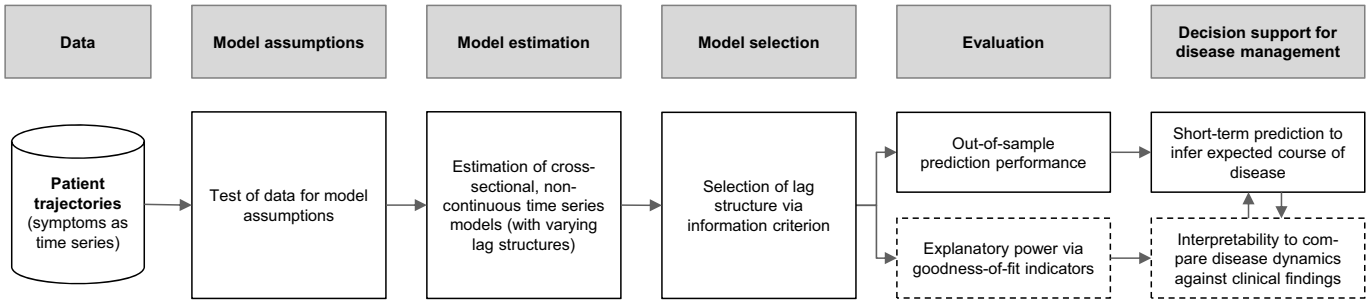

Fig. 2. Research framework for evaluating our cross-sectional time series model, so that we estimate parameters to *multiple* time series within a cohort (as opposed to only a *single* time series as in traditional time series analysis) for yielding an interpretable prediction of disease progression.

then informs health professionals that are supposed to adapt their decision-making accordingly. To this end, these models predict individual events and thus differ from the objective of our study, that is, they are not designed to model the actual course of a disease via a cross-sectional time series model as in this work.

Second, optimizing the design of treatment plans when the treatment plan has already been determined. This involves decision-making along multiple dimensions, namely the timing of medication (e.g., [15]), its type (e.g., [8,9]), and the dosage (e.g., [47]). The underlying decision-making problem is often formalized via a (partially-observable) Markov decision process (e.g., [15]) or, when measuring the responsiveness of patients to treatments, a bandit model (e.g., [47]). However, these models address only the design of a given treatment plan that should be optimized; they do not provide decision support regarding the choice among multiple treatment plans.

Third, prediction models for disease progressions infer the expected course of the trajectory in the short run, so that uncertainties regarding the progression are alleviated [60]. This is supposed to inform practitioners and patients, who can then consider the expected course in their decision-making [18]. A common approach builds upon patient-individual models, whereby a time series model is fitted to the trajectory of a single patient (e.g., [24]). However, patient-individual time series models are subject to limitations. On the one hand, the models can leverage the observed disease dynamics from a complete patient cohort. On the other hand, the models can only make predictions when observations from the past time series of the patient are already known (i.e., model parameters must first be estimated, thereby prohibiting offline settings where a model is directly applied to an unseen patient). As a remedy, previous research has developed various machine learning approaches. For example, Wang et al. [62] develop a high-order multi-task learning model to predict the cognitive performance from phenotypic markers in patients with Alzheimer's disease. Gaussian-process-based models have been proposed for predicting disease trajectories in chronic kidney disease from biomarkers [26]. Schulam and Saria [58] develop a Gaussian-process-based model to predict the course of autoimmune diseases from its previous trajectory. The aforementioned works are based on machine learning models that largely act in black-box fashion. Owing to this, interpretability with respect to the inferences is largely precluded.

Altogether, there is a scarcity of interpretable time series models that infer the expected course of diseases. Our work is therefore inspired by other decision-making fields, such as finance, energy production, or operations, where many decisions are based upon short-term expectations. Here, linear autoregressive models are successfully applied to predict the development of various performance metrics (e.g., [1,14,48,51]). The drawback is that these models fit one model to one time series, whereas our objective is to predict disease trajectories from generalizing over a sample of different patient time series. To this end, this work develops a cross-sectional time series model that fulfills key requirements from health practice: (1) It allows for a high degree of interpretability due to a parsimonious model specification. This is in line with clinical practice, where it is widely argued that interpretability is a key prerequisite for data-driven models in healthcare analytics [54]. As a result, our parsimonious model specification ensures accountability. (2) Our model is cross-sectional; that is, it is designed for cohort-wide studies. As a result, inferences are made from *one* model based on *multiple* time series within a patient cohort (rather than fitting only *one* model to the time series of only *one* patient). (3) We follow a theory-informed approach by modeling the intrapersonal variability in the disease progression via an autoregression analogous to earlier findings from clinical research [46].

3. Model development

This section develops our cross-sectional time series models. The models yield cohort-wide parameters estimated based on data from *multiple* patients (rather than a single, patient-individual time series). For this purpose, we follow the research framework in Fig. 2.

3.1. Problem statement

Our research objective is to model the expected course of a disease based on data from longitudinal monitoring. Formally, the longitudinal input is given by symptoms X that have been recorded for a patient cohort $i=1,\ldots,N$. The symptoms $X\in\mathbb{R}\cup\{\_\}$ are collected across equally-spaced time steps $t=1,\ldots,T$. Furthermore, the symptoms are described by either a numeric value or missing at random (denoted by "_"). Due to the latter, the time series of symptoms for an individual patient comprises multiple segments that are indexed by s. Subsequently, the overall input is formally given by non-contiguous segments X_{ist} for patients $i=1,\ldots,N$, contiguous segments $s=1,\ldots,\sigma_i$, and corresponding time steps $t=1,\ldots,\tau_{is}$. Based on this, our objective is to infer the expected symptoms in the next step, that is, $E[X_{i,\,s,\,t+1}]$. This prediction must fulfill additional requirements from clinical practice as listed below.

First, our model should provide the basis for decision-making in clinical practice and, hence, must be interpretable. If a treatment decision is supposed to be based on the expected progression of a condition, the health professional needs to understand which dynamics of the disease progression have led to this inference. This is demanded by health professionals in practice, so that they can understand how inferences are formed. Because of this, interpretability was considered a key requirement for data-driven modeling that involves high-stakes decisions such as in healthcare [28,54], where trust is a basis for acceptance [28,59]. Hence, the expected progression is modeled via a linear dependency.

Second, our settings demands time series models that are cross-sectional. Accordingly, we need to fit *one* model to *multiple* time series (i.e., from the different patients $i=1,\ldots,N$ in our cohort). This differs from traditional time series models that fit *one* model to *one* time series. Formally, the parameters in our works should be estimated via $E[X_{i,\,s,\,t+1} \mid X_{ist}\ \forall\ i]$ instead of $E[X_{i,\,s,\,t+1} \mid X_{ist}$ for $i = i^*]$. While cross-sectional

time series models have been developed earlier [56], our work demands a few customizations as discussed later.

Third, our model development must closely adhere to prior findings from clinical research. To this end, we choose time series models that model the intrapersonal variability of symptoms through autoregression, i.e., where future symptoms depends on the current state of symptoms. Autoregression has been observed in the progression of various symptoms, such as, for instance, pain [46]. By formalizing autoregression in our model, we follow a theory-informed approach.

There is a wide range of models that can theoretically perform short-term predictions from univariate time series, as can be seen from other decision problems. These include, for instance, Gaussian-process-based models (e.g., [58]), support vector machines (e.g., [37]), and neural networks (e.g., [53]). However, most of the technically suitable modeling approaches are precluded as they violate our first premise; namely, interpretability for practitioners. Yet there is one very common model class that has a rich history of short-term prediction of time series with a high amount of variability, while yet offering a parsimonious, linear configuration. ARMA models are still widely regarded as state of the art for many forecasting problems where the underlying course follows autoregressive dynamics (e.g., [51]). Their main advantages lie in their simplicity, and their ability to "capture complex patterns of temporal correlation" [16]. That is why, in order to accommodate the above requirements, we propose the use of a cross-sectional, non-contiguous ARMA(1,1) model. Here, the parsimonious lag structure allows health professionals to infer the tendency of a condition to persist or mean reverse, thereby driving the decision-making behind care (e.g., [25]). The aforementioned ARMA(1,1) is designed in a way that it incorporates a time series with missing values and that thus comprises of multiple segments s.

The above cross-sectional ARMA(1,1) model presents our baseline model. It is later compared with alternative lag structures. Thereby, we can quantify the trade-off between a potential improvement in predictive power but at the cost of limited interpretability. Furthermore, we expect that symptoms might be characterized by temporary phases with larger variability (e.g., [21,41,42]). Therefore, the proposed ARMA model can be extended by a cross-sectional GARCH process for the purpose of modeling volatility clusters in the time series. The GARCH model describes the volatility of the residuals from the ARMA model when they follow a conditional structure with clusters of temporarily stronger variability. To the best of our knowledge, there is a scarcity of cross-sectional GARCH models in prior literature and, therefore, such a model represents another contribution of this work.

3.2. Cross-sectional time series models

The estimation procedure for the cross-sectional, non-contiguous time series model is as follows.

3.2.1. Cross-sectional ARMA model

The cross-sectional, non-contiguous ARMA(p,q) model with p autoregressive and q moving-average terms is defined by

$$X_{ist} = \alpha + \sum_{z=1}^{p} \beta_z X_{i,s,t-z} + \sum_{z=1}^{q} \gamma_z \varepsilon_{i,s,t-z} + \varepsilon_{ist}, \quad \text{for all } i, s, \text{ and } t, \tag{1}$$

with intercept α, and coefficients $\beta_1, \ldots, \beta_p$ for the autoregressive terms, and coefficients $\gamma_1, \ldots, \gamma_q$ for the moving-average terms. The difference to a traditional ARMA model [32] lies in the model parameters: these are estimated with data from patients $i = 1,\ldots, N$ (i.e., cross-sectional) and, furthermore, each time series is composed of different segments $s = 1,\ldots, \sigma_i$, which we refer to as "non-contiguous".

The parameters of the cross-sectional, non-contiguous ARMA model are determined via maximum likelihood estimation (MLE) as follows. For this purpose, let us denote the parameter configuration of the model by $\Theta_{\text{ARMA}} = (\alpha,\beta_1,\ldots,\beta_p,\gamma_1,\ldots,\gamma_q)$. Accordingly, the overall log-likelihood L of a cross-sectional, non-contiguous ARMA model is maximized in order to identify the optimal parameter configuration Θ_{ARMA}^*, i.e., $\Theta_{\text{ARMA}}^* = \underset{\Theta}{\text{argmax}} L(\Theta_{\text{ARMA}})$. Then, the log-likelihood is given via

$$L(\Theta_{\text{ARMA}}) = \frac{\sum_{i=1}^{N}\sum_{s=1}^{\sigma_i}\sum_{t=1}^{\tau_{is}} l_{ist}(\Theta_{\text{ARMA}})}{\sum_{i=1}^{N}\sum_{s=1}^{\sigma_i}\tau_{is}} \quad \text{with} \quad l_{ist}(\Theta_{\text{ARMA}}) = -\frac{1}{2}\log(\sigma^2) - \frac{\varepsilon_{ist}^2}{2\sigma^2} \tag{2}$$

where σ^2 denotes the variance across all observations X_{ist}.

The above model is subject to several assumptions. First, it assumes that the underlying time series is stationary. This is tested later via unit-root-test according from Levin et al. [44] that has been specifically developed for cross-sectional time series data (cf. [35]). Assuming stationarity is also in line with theory from clinical research. According to the Corbin-Strauss trajectory framework, many long-lasting and, especially, chronic diseases have no cure; instead, the underlying symptoms are stabilized [17]. This should thereby yield stationarity. Second, the estimation must cater for autocorrelation, heteroscedasticity, and potential cross-sectional correlation inferences. In our work, these are corrected according to Driscoll and Kraay [22], which represents a reliable measure for robust standard errors in cross-sectional settings.

3.2.2. Cross-sectional GARCH model

As part of the model extensions later, the cross-sectional, non-contiguous ARMA model is augmented by a GARCH part in order to consider volatility in the errors of the ARMA. This results in a cross-sectional, non-contiguous ARMA-GARCH(η,ζ) model with η ARCH terms and ζ GARCH terms.

Accordingly, a cross-sectional, non-contiguous ARMA(p,q) model is first fitted, yielding the residuals ε_{ist}. Let us further introduce h_{ist}, which refers to the conditional variance of the residuals. Then, the GARCH (η,ζ) model consists of the ARCH process (modeling a linear dependency between residuals and conditional variance) and the GARCH process (modeling autoregression of the conditional variance). Mathematically, this is given by

$$h_{ist} = \phi + \sum_{z=1}^{\eta} \psi_z \varepsilon_{i,j,t-z}^2 + \sum_{z=1}^{\zeta} \rho_z h_{i,j,t-z} + \varepsilon_{ist}, \tag{3}$$

with intercept ϕ, coefficients $\psi_1, \ldots, \psi_\eta$ for the ARCH part, coefficients $\rho_1, \ldots, \rho_\zeta$ for the GARCH part, and, furthermore, residuals ε_{ist}.

Again, the parameters are determined via MLE. As such, let $\Theta_{\text{GARCH}} = (\phi,\psi_1,\ldots,\psi_\eta,\rho_1,\ldots,\rho_\zeta)$ define the parameter configuration of the cross-sectional GARCH model. Based on this, the overall log-likelihood L is maximized, i.e., $\Theta_{\text{GARCH}}^* = \underset{\Theta}{\text{argmax}} L(\Theta_{\text{GARCH}})$. Both the cross-sectional and non-contiguous nature of the data must be considered. For this reason, the log-likelihood from Bollerslev [10] is adapted by a summation across patients and time series segments. This results into

$$L(\Theta_{\text{GARCH}}) = \frac{\sum_{i=1}^{N}\sum_{s=1}^{\sigma_i}\sum_{t=1}^{\tau_{is}} l_{ist}(\Theta_{\text{GARCH}})}{\sum_{i=1}^{N}\sum_{s=1}^{\sigma_i}\tau_{is}} \quad \text{with} \quad l_{ist}(\Theta_{\text{GARCH}}) = -\frac{1}{2}\log(h_{ist}) - \frac{\varepsilon_{ist}^2}{2h_{ist}}. \tag{4}$$

3.3. Model selection

As a default, an ARMA(1,1) model is considered based on the requirements of this study, which include interpretability. In addition, it is compared against alternative model specifications, namely ARMA(p,q) models with different lag structures. Here, the comparison

is based on the corrected Akaike information criterion, AICc for short [11].

The AICc represents a common metric for performing model selection of ARMA models [38,63] and entails advantages for our study: the AICc particularly penalizes the number of predictors in order to reduce the risk of overfitting. It thereby offers a correction for small sample sizes where the number of time steps is not considerably larger than k^2 (i. e., the squared number of parameters), while not holding any disadvantage towards the uncorrected AIC for larger samples [63]. This is specifically useful for the study design in this work where time series comprise of short segments (with length $\tau_i < 100$). The AICc is adjusted to sum up across all patients in order to serve a cross-sectional setting.[1] This yields

$$\text{AICc} = \frac{1}{N}\sum_{i=1}^{N}(2k - 2L_i + c_i) \quad \text{with} \quad c_i = \frac{2k(k+1)}{\tau_i - k - 1}, \tag{5}$$

where c_i specifies a correction factor for small samples and where $k = p + q + 1$ is the number of coefficients in the ARMA model (or GARCH model, respectively). The variable L_i denotes the log-likelihood as defined earlier for ARMA or GARCH models, respectively, but it is here applied to each patient i, thus summing up across all respective time series segments σ_i.

For comparability, the Bayesian information criterion (BIC) is also reported. The BIC is known to perform well for moderate sample sizes [63]. It is given by

$$\text{BIC} = \frac{1}{N}\sum_{i=1}^{N}(k\log(\tau_i) - 2L_i). \tag{6}$$

3.4. Evaluation of prediction performance

The prediction performance is measured by comparing the model against ground-truth in out-of-sample observations.

The procedure follows the conventional approach in predictive modeling [33]. That is, the original dataset is split into two disjunct subsets $\mathscr{X}$ and $\mathscr{Y}$ that serve for training and testing, respectively. The split ratio is determined by analyzing corresponding mean squared error (MSE) curves in accordance with Dobbin and Simon [20]. The training set is utilized to calibrate the model parameters by means of MLE. In a subsequent step, the fitted model is applied to the testing set where predictions can be made on unseen observations. The deviation is then recorded by the root mean squared error (RMSE) and the mean absolute error (MAE). To provide a measure for prediction uncertainty, prediction intervals at the 80% and 95% level are calculated. As residuals are not expected to be normally distributed, prediction intervals are calculated by utilizing a bootstrapping resampling method according to Thombs and Schucany [61].

The prediction power of our cross-sectional, non-contiguous time series model is evaluated based on the following baselines: (a) the in-sample mean, (b) the first-order lag as a trivial prediction, and (c) patient-individual models:

(a) The first baseline is given by $X_{ist} = \overline{\mathscr{X}}$, where $\overline{\mathscr{X}}$ refers to the in-sample mean. Hence, these baselines measure the prognostic capacity of our cross-sectional model over a global average.
(b) The first-order lag represents a trivial prediction of perfect autocorrelation, where the expected observation, X_{ist}, is set to the previous observation, $X_{i,\ s,\ t-1}$. Here, the assumption is that the symptom remains constant in consecutive time steps.
(c) Patient-individual models refer to ARMA models that were estimated based on the historic time series from an individual patient. In order to predict X_{ist}, an ARMA$_i(p,q)$ model is estimated based on observations across time steps 1, …, $t-1$ from patient i. Therefore, all observations except the current one are used and, as a result, different parameters are estimated for each t. Here, we reiterate that such models are still non-contiguous in the sense that they generalize over the different segments s of each patient in order to address potentially missing observations (i.e., formally, a cross-sectional, non-contiguous model is fitted but is set to $N = 1$). Altogether, this baseline might benefit from parameters that can be personalized to patient-specific disease dynamics but, since it cannot generalize to cohorts, its performance in offline settings (i.e., cold starts where a past time series is largely absent) is impeded.

3.5. Evaluation of explanatory performance

The explanatory performance is measured via the goodness-of-fit, i.e., R^2. In the context of this work, the R^2 quantifies how well the model can capture intrapersonal variability of symptoms beyond the average severity $\overline{X}$ across all observations X_{ist}. Formally, the R^2 is adapted to the cross-sectional, non-contiguous time series setting via

$$R^2 = \frac{1}{N}\sum_{i=1}^{N}\left[1 - \frac{\sum_{s=1}^{\sigma_i}\sum_{t=1}^{\tau_{is}}\varepsilon_{ist}^2}{\sum_{s=1}^{\sigma_i}\sum_{t=1}^{\tau_{is}}(X_{ist} - \overline{X})^2}\right]. \tag{7}$$

The explanatory power of our cross-sectional, non-contiguous time series model is again evaluated against the following baselines: (a) an intercept-only model and (b) patient-individual models:

(a) The intercept-only model is given by $\overline{\mathscr{X}}$, which here refers to the sample mean. Hence, this baseline measures the explanatory power against common practice for making inferences in clinical settings [46].
(b) Patient-individual models are used to evaluate the trade-off between allowing for larger flexibility versus maintaining interpretability through fewer parameters. Formally, the patient-individual models estimate a separate ARMA$_i$(1,1) to the time series of each patient i. Therefore, the parameters can vary across patients $i = 1$, …, N.

4. Empirical setting: low back pain

Our cross-sectional time series model is evaluated based on a longitudinal study of patients suffering from low back pain. This condition was specifically chosen for the following reasons. First, it ranks among the most prevalent chronic diseases around the world, resulting in the highest rates of YLD globally [39]. Second, treatment planning for non-specific low back pain should be primarily based on the expected course rather than a (bio-)medical diagnosis [18,25]. Third, low back pain reveals considerable intrapersonal variability [31], thus rendering accurate predictions of the progression challenging. Finally, predictions could also help patients adapt their daily activities, according to their individual circumstances.

Our study comprises 928 patients with low back pain; see Nielsen et al. [50] for detail on the inclusion criteria of our study. Specifically, patients were required to be between 18 and 65 years old and needed to be able to respond to longitudinal monitoring. The data collection consists of two parts (see Fig. 3): (1) the longitudinal pain data and (2) a baseline questionnaire. Both are detailed below.

The longitudinal pain data set consists of patients' weekly pain levels over the course of 52 weeks. For this purpose, patients were asked to report their current pain intensity [41] according to the so-called numerical rating scale (NRS) as a quasi-standard in medical studies

[1] We validate the extension of AICc and later BIC to cross-sectional settings with a simulated dataset that is based on autoregressive models of a predefined lag order. Both metrics perform as described by Hurvich and Tsai [38] and Wong and Li [63].

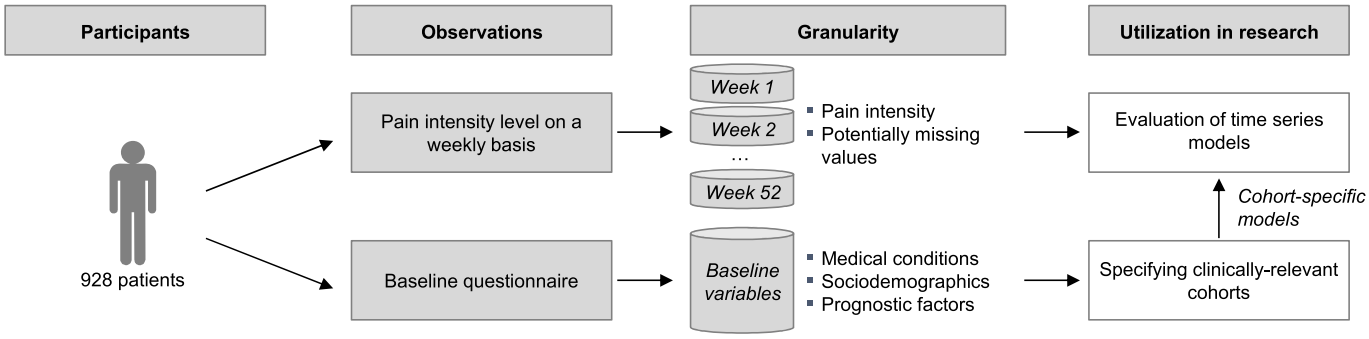

Fig. 3. The data set was obtained from a longitudinal study, consisting of 928 patients suffering from low back pain. For each patient, health-related information was collected based on (1) longitudinal pain data in weekly resolution and (2) a baseline questionnaire asking for variables specifying clinically-relevant cohorts.

Table 1
Panel A lists the input for the cross-sectional time series models. Panel B reports the baseline variables that are used as part of the cohort-specific sensitivity analyses.

Name	Description	Values	Source
Panel A: observations for time series models			
Pain intensity	Pain intensity according to NRS scale [55]	0–10	Weekly report
Panel B: baseline variables for cohort-specific sensitivity analyses			
Sex	Sex of patient	Female, male	Baseline questionnaire
Physical workload	Typical physical activity at work	Mostly sitting, mostly physical	Baseline questionnaire
Smoking habit	Smoking habit of patient	Smoker, non-smoker	Baseline questionnaire
Leg pain	Pain in legs in addition to low back pain	Yes, no	Baseline questionnaire
Comorbidities	Patient suffers from other chronic disease	Yes, no	Baseline questionnaire

[55]. The scale ranges from 0 (i.e., no pain) to 10 (i.e., worst imaginable pain).

The baseline questionnaire comprises variables regarding the medical condition, as well as prognostic factors (e.g., comorbidities), and sociodemographic variables [50]; see Table 1 for details. These variables are later used as part of a sensitivity analysis in which the model is re-estimated for specific subgroups. This allows us to perform subgrouping, where the models are personalized to patients with a similar risk profile.

Summary statistics are as follows. Of the 928 participating patients, 19 did not report any longitudinal data and were therefore excluded for further analysis in this paper. For the 909 remaining patients, the mean pain level amounts to 1.65 with a standard deviation of 1.72. On average, 52% of weekly pain levels are reported as 0 (i.e., no pain), while 13% of weekly pain levels have not been reported. The median contiguous length of pain recordings (i.e., without missing values) consists of 12 values. Subsequently, this yields non-contiguous patient time series, which is addressed by the proposed model specification.

The model assumptions have been confirmed as follows: The cross-sectional time series data from the patient cohort has been tested for stationarity with the unit-root test formulated by Levin et al. [44]. The test rejects the null hypothesis at a significance level of 0.001 and, hence, stationarity should be assumed. This is in line with clinical research, according to which, non-specific low back pain is oftentimes chronic [25,31]. Therefore, the objective for health professionals is not to find a cure but to stabilize symptoms over time [17].

Table 2
As part of the model selection, a variety of cross-sectional ARMA models are compared based on AICc and BIC. Here, results are listed for lag structures with $p + q \leq 5$ (panel A; top-3 models) and, as comparison, easily interpretable models (panel B). Preferred values for information criteria are in bold font.

Model specification		Information criteria		Goodness-of-fit (R^2)
p	q	AICc	BIC	
Panel A: best-performing models				
1	1	**157.87**	**167.18**	**0.53**
1	0	162.14	169.75	0.48
2	1	159.45	170.36	0.53
Panel B: comparison against interpretable benchmarks				
1	0	162.14	169.75	0.48
0	1	186.95	194.56	0.31
Intercept-only model		205.72	211.55	–

5. Results

This section evaluates the cross-sectional, non-contiguous time series models with data from our study of low back pain as follows. First, a model selection is performed, which confirms that the proposed cross-sectional ARMA(1,1) model attains a superior fit. Second, this model is used for a detailed evaluation of the prediction performance and, third, the estimation results are reported in order to demonstrate interpretability. Later, the model is subject to extensions.

5.1. Model selection

Model selection is performed by estimating a variety of models with different lag structures of up to 5 lags (i.e., $p + q \leq 5$). These are then compared based on AICc and BIC. The results for the top-3 models are reported in Table 2. Both information criteria favor the choice of a cross-sectional ARMA(1,1) model with one autoregressive and one moving-average term.[2] Notably, this model outperforms models with higher-order lag structures. This is advantageous in our case as it yields a parsimonious model specification and thus ensures interpretability. Lastly, the information criteria from the intercept-only are reduced by a

[2] We perform a series of model checks. The model residuals are still subject to a remaining but slight serial correlation. The Ljung-Box test statistic amounts to 3.90, which corresponds to P-value of 0.05. Furthermore, the Breusch-Pagan test suggests the presence of heteroskedasticity with a test statistic that is significant at the 0.01 level. An additional Pesaran test also suggests cross-sectional dependency with a test statistics at the 0.001 level. Accordingly, we later compute robust standard errors following Driscoll-Kraay that correct for cross-sectional correlation, serial correlation, and heteroskedasticity [22].

Table 3
Reported are the out-of-sample prediction performance for point forecasts and 80% and 95% prediction intervals of the cross-sectional ARMA(1,1) model. Prediction intervals are detailed as average deviation from point forecasts. Performance is compared against several baselines: an intercept-only model, a first-order lag as a trivial prediction, and patient-individual models.

Type	Model	Point forecast accuracy		Prediction interval			
		RMSE	MAE	Lower 80%	Upper 80%	Lower 95%	Upper 95%
Baselines	Intercept-only model	2.18	1.75	−1.46	+3.54	−1.46	+5.54
	First-order lag	1.60	0.75	−2.00	+1.00	−4.00	+4.00
	Patient-individual models	1.48	0.93	−0.97	+1.87	−2.53	+4.02
Proposed model	Cross-sectional ARMA(1,1)	1.35	0.83	−1.54	+1.16	−2.93	+3.23

comparably large margin, thus pointing out that a global average is not sufficient for describing the variability in pain levels. Altogether, the cross-sectional ARMA(1,1) model provides the basis for all subsequent evaluations.

5.2. Prediction performance of disease progression

The prime objective is to model the expected course of the disease. Hence, the selected ARMA(1,1) model is compared with different baselines as detailed in Table 3. For this, the parameters of the cross-sectional ARMA(1,1) model are estimated on the training set, while Table 3 reports the out-of-sample prediction performance for the one-week-ahead pain levels on the test set (i.e., on unseen patients). As part of a robustness check, we also computed the MSE as a function of the split ratio. We found that, as a result of the parsimonious model specification, the results remain stable. This gives us the freedom to choose a test set size at the larger end of commonly used ratios to validate reliability of the out-of-sample performance evaluations [20] as more patients are included therein. Accordingly, we assign 40% for training and 60% for testing. The following observations can be made.

First, the cross-sectional, non-contiguous ARMA(1,1) model attains an out-of-sample performance that is comparable to the in-sample performance. Specifically, it captures 53% of variance in out-of-sample settings. This equals the in-sample R^2. As performance stays steady over varying test and training sample splits, that means that only a few random patient samples are required to estimate a global model with our proposed parsimonious configuration. Second, the cross-sectional ARMA(1,1) model performs better than all baselines, namely (1) an intercept-only model, (2) a first-order lag as a trivial prediction, and (3) patient-individual ARMA models. The intercept-only model still finds broad application in medical practice [46], yet it is outperformed. Specifically, replacing the intercept-only model with our proposed model reduces the RMSE from 2.18 to 1.35, which is a reduction by 38%. Similarly, the first-order lag is not on par with our model, thus pointing out that mean-reversion is important for modeling the intrapersonal variability of the disease. For instance, the first-order lag only captures 25% of the out-of-sample variability in pain as compared to 53% for our cross-sectional ARMA(1,1). In sum, the results confirm that the intrapersonal variability in the progression of pain can be effectively modeled via autoregression. Accordingly, the previously-experienced pain level carries predictive value for future pain. Third, the prediction intervals[3] further show that our cross-sectional ARMA(1,1) model narrows the corridor of expected disease developments, compared to all baselines. The prediction interval indicates that 80% of future pain levels are within the average deviation corridor of (−1.54; +1.16) around the point forecast.

As an additional observation from the analysis, the patient-individual autoregressive models have challenges in providing accurate predictions. This is line with our expectations: Such models require sufficient data for *each* patient in order to be able to estimate parameters. To provide further insights, we performed additional checks where we compared the RMSE as a function of the available data. In fact, the RMSE stabilizes only when at least 19 weeks of observations are available for a patient. In contrast, cross-sectional time series models generalize over cohorts and thus can be directly utilized in offline settings, that is, when making inferences for unseen patients.

Stated are RMSE and prediction intervals from resampling test sets and bootstrapping respective residuals. MAE is given as an additional metric for reasons of interpetability. Note that all models have been estimated via an L2-norm and, hence, the RMSE should be used for comparing model fit.

5.3. Estimation results

In Table 4, the estimation results for the cross-sectional ARMA(1,1) are reported, thereby demonstrating that the proposed model is interpretable. Evidently, the autoregressive term is associated with a coefficient of $\beta_1 = 0.942$ (P-value < 0.001). Given the size of the coefficient, the subsequent pain levels are described by the pain level from the previous week to a large extent. This is analogous to previous studies in medical research, according to which, the progression of symptoms of long-lasting diseases is subject to autoregressive dynamics (e.g., [42,46]). Our proposed cross-sectional model confirms this tendency for low back pain. Accordingly, the strong autoregressive term with an amplitude close to 1 indicates persistence of a condition with slow recovery from one week to another. Furthermore, the moving-average term with a coefficient of $\gamma_1 = -0.492$ (P-value < 0.001) is statistically significant. This implies mean-reversion or stabilization at preceding levels, whereby a sudden pain peak should largely revert afterwards.

In order to shed further light, the actual course of the disease is compared to the fitted model for two example patients; Fig. 4 details the time series with observed pain from two different patients with the estimated values from the proposed cross-sectional ARMA(1,1) model. Here, autoregression can be observed in many phases of both time series: Oftentimes pain levels remained fairly consistent, so that there was hardly any change from one week to another (or there was only a minor fluctuation around a base level). Both time series reveal a partial reversal after a pain episode (i.e., a sudden period of stronger pain or temporary recovery). The first patient in Fig. 4 depicts such dynamics, for instance, in weeks 1 to 7 where pain after smaller shocks returns to a (slightly) lower level and, furthermore, in weeks 36 to 42 where pain intensity quickly drops and then largely reverts again. The second patient has such dynamics throughout the complete course. The patient experiences reversal after smaller fluctuations and a quick recovery after a strong pain peak in week 17.

5.4. Sensitivity analysis with cohort-specific estimation results

Disease dynamics are likely to vary across subgroups. In order to substantiate this with empirical evidence, a sensitivity analysis is

[3] Prediction intervals were computed from bootstrapped residuals as residuals are not normally distributed. For comparison, the trivial prediction interval for our cross-sectional ARMA(1,1) model would be (−1.74; +1.74) at the 80% level and (−2.66; +2.66) at the 95% level.

Table 4
Parameter estimates (on the full sample) for the cross-sectional ARMA(1,1) model.

Parameter	Estimate
Intercept (α)	0.094* (0.041)
Autoregressive term (β_1)	0.942*** (0.061)
Moving-average term (γ_1)	−0.492*** (0.073)

Stated: coefficients with Driscoll-Kraay-corrected standard errors in parentheses; significance levels are *** 0.001, ** 0.01, * 0.05.

performed in which cohort-specific estimation results are reported. This highlights a strength of our approach, as it can be used for the inference of tailored results for specific cohorts.

To this end, Table 5 compares the model coefficients across subgroups, namely the duration after disease onset, sex, physical workload, smoking habit, additional leg pain (i.e., as an example for prognostic factor from clinical research that is used to describe risk profiles), and cohorts with/without comorbidities. Overall, the disease dynamics are fairly similar across subgroups. In a later stage in the course of the disease (e.g., six months after initial diagnosis), the progression is slightly more persistent or self-stabilizing, which is indicated by a more negative moving-average term. The higher autoregressive term also hints towards more stability, yet the difference is not statistically significant. Overall, this investigation confirms results from previous studies that have pointed out the difficulty to determine prognostic factors characterizing disease profiles for personalization [41,49].

Table 6
This table compares the proposed cross-sectional ARMA(1,1) model against alternative model specifications, namely a variant with random effects (i.e., patient-individual estimates of the intercept), patient-individual ARMA(1,1) models, and an intercept-only model (i.e., population-wide estimates). The table also lists the variance σ^2 of the performance metrics (i.e., R^2, RMSE, MAE) when computed at the level of patients.

Model	R^2	$\sigma^2(R^2)$	RMSE	σ^2(RMSE)	MAE	σ^2(MAE)
Cross-sectional ARMA(1,1)	0.53	0.09	1.35	0.33	0.82	0.22
Random-effects cross-sectional ARMA(1,1)	0.55	0.08	1.27	0.33	0.77	0.26
Patient-individual ARMA(1,1)	0.60	0.07	1.22	0.27	0.74	0.21
Intercept-only model	–	–	2.18	0.85	1.75	0.73

5.5. Robustness checks

As part of the robustness checks, the cross-sectional, non-contiguous ARMA model is compared against alternative approaches for modeling the expected course of the disease. Here, the alternative model specifications include a variant of the cross-sectional ARMA(1,1) with random effects (i.e., patient-individual estimates of the intercept), patient-individual ARMA(1,1) models, and an intercept-only model (i.e.,

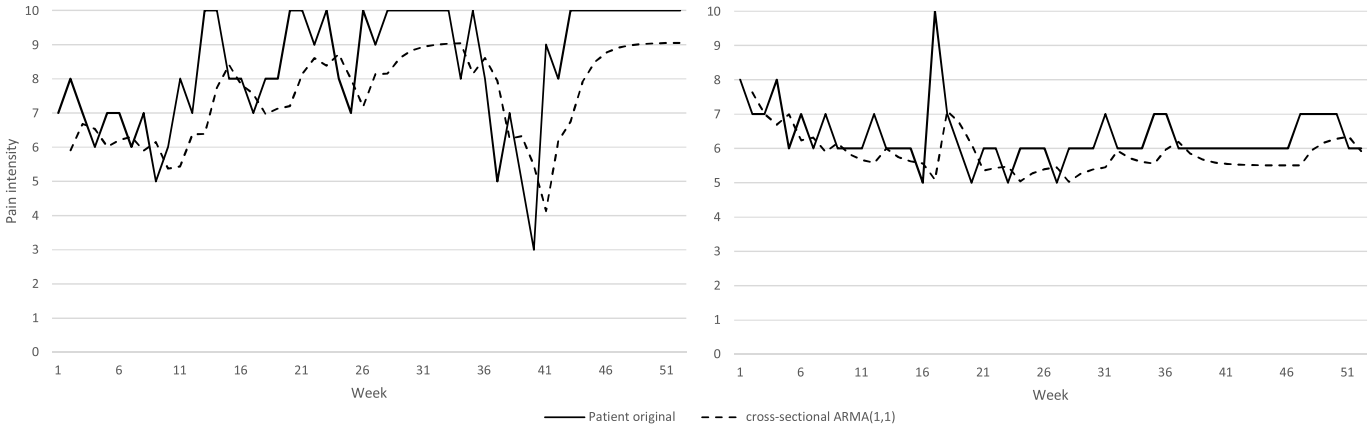

Fig. 4. The plot shows the observed pain level over time for two example patients and, in comparison, the fitted values from the cross-sectional ARMA(1,1) model.

Table 5
Comparison of estimated parameters for cross-sectional ARMA(1,1) models across different cohorts. Reported are the respective coefficients, the difference across subgroups, and the z-values when testing the difference for significance.

Parameter	All patients	By time		By sex		By physical workload		By smoking habit		By leg pain		By comorbidities	
		First 6 months	Last 6 months	Female	Male	Sitting	Physical	Smoker	Non-smoker	Yes	No	Yes	No
Intercept (α)	0.094	0.172	0.045	0.118	0.081	0.088	0.116	0.095	0.103	0.097	0.106	0.113	0.090
		−0.127		−0.036		+0.028		+0.009		+0.009		−0.024	
		(−1.015)		(−0.592)		(0.371)		(0.113)		(0.134)		(−0.363)	
Autoregressive term (β_1)	0.942	0.915	0.961	0.932	0.951	0.946	0.937	0.954	0.939	0.951	0.921	0.945	0.939
		+0.047		+0.019		−0.009		−0.015		−0.030		−0.006	
		(0.601)		(0.219)		(−0.103)		(−0.178)		(−0.333)		(−0.073)	
Moving-average term (γ_1)	−0.492	−0.394	−0.606	−0.475	−0.509	−0.508	−0.479	−0.483	−0.495	−0.505	−0.48	−0.485	−0.497
		−0.213*		−0.033		+0.029		−0.013		+0.026		−0.012	
		(−2.189)		(−0.314)		(0.262)		(−0.106)		(0.237)		(−0.113)	

Stated: coefficients and their deltas between specific subgroups; z-values in parentheses with Driscoll-Kraay correction; significance levels are *** 0.001, ** 0.01, * 0.05.

population-wide estimates). This allows us to attribute the difference in performance when using a cohort-wide model with a parsimonious specification, instead of modeling the between-patient heterogeneity.

Table 6 lists the in-sample results. First, the between-patient variance in the baseline pain level is only marginally helpful in improving fit. This is shown by an R^2 of 0.55 from the random effects ARMA as compared to 0.53 for our proposed cross-sectional ARMA.[4] Second, models with patient-individual dynamics should, in theory, have larger flexibility to capturing the between-patient heterogeneity better than a cohort-wide model. This is indicated by an $R^2 = 0.60$ of patient-individual ARMA(1,1) models. The larger flexibility, however, also leads to the disadvantage of greatly increasing the number of parameters: Patient-individual ARMA(1,1) models entail $N(p + q + 1) = 2727$ parameters, whereas this is reduced to a $p + q + 1 = 3$ parameters for our cross-sectional model in order to achieve cohort-wide interpretability. Overall, we find that the patient-individual ARMA(1,1) models improve the explained variance by 0.07. However, the applicability of such models in out-of-sample settings is impeded, since, for unseen patients, data on the disease progression must be collected before being able to perform an estimation. According to Table 6, the between-patient variance of performance measures is fairly stable. This confirms that the proposed model can generalize well across patient cohorts.

5.6. Extension to cross-sectional ARMA-GARCH model

The progression of pain might be subject to volatility clusters and, therefore, the cross-sectional ARMA(1,1) model is extended by a cross-sectional GARCH model. This allows us to model clusters where pain is temporarily subject to strong fluctuations, rather than being described by a constant variance. Table 7 details the corresponding estimation results when estimating a cross-sectional GARCH model to the residuals of the above cross-sectional ARMA(1,1) model. Here, the choice of a GARCH(2,0) model is preferred. As a result, there is a tendency that volatility clusters for pain episodes persist for at least two consecutive weeks. According to the medical researchers of our author team, this is an interesting insight that helps accumulating knowledge on intrapersonal variability of pain over the existing works (cf. [46]).

Table 8 reports the estimated parameters from the selected GARCH (2,0) model. The ARCH parameters measure the extent to which a volatility shock (i.e., sudden fluctuation) in form of the error of the ARMA model is passed on to the next weeks' volatility. Both coefficients, $\psi_1 = 0.35$ and $\psi_2 = 0.31$, indicate that the onset of an unpredictable volatility shock affects the fluctuation of pain in the following two weeks: A shock diminishes only slowly in the following weeks, yet it does not persist. Hence, volatility tends to manifest in clusters.

Second-order ARCH parameters with $\psi_1, \psi_2 > 0.30$ indicate volatile spikes that pass on to the following two weeks and, therefore, link to the uncertainty in the predictions. Hence, this differs from unconditional variance or persistent volatility shocks which would be indicated by high GARCH parameters [14]. However, the observed volatility for patients of low back pain seems to be of a rather short episodic nature. In other words, volatility tends to manifest in clusters. Therefore, we follow the terminology from the literature [12] and describe such trajectories (i.e., with ARCH Term > 0.01 and low GARCH term) as particularly “spiky”. Previous work defined fluctuations as deviations from the mean as a given property for each time point [40], while this work adds a quantification of the extent to which it carries through multiple consecutive time points.

[4] As an additional robustness check, we estimated a multilevel model (i.e., so-called Arfima-MLM for repeated cross-sectional data) according to Lebo and Weber [43]. This model fits an aggregate linear regressor dependent on time that is being corrected by an aggregate ARMA process and patient-individual random effects. The random effects should be beneficial to model the between-patient heterogeneity in explanatory settings. As a result, the model achieves a MAE of 0.62 and a RMSE of 0.99 in an in-sample setting. However, the model is designed for explanatory purposes and not prediction (the reason is that, in the prediction, the mean value of the random effects must be used as a best guess, since the true value is still unknown)

6. Discussion

6.1. Implications for practice

Many long-lasting diseases, such as low back pain, are characterized by a course that is subject to considerable variability. Accordingly, it makes a decisive difference for treatment planning if it is known whether a symptom is expected to decline or fluctuate, or whether it is expected to persist (e.g., [25]). To ensure effective disease management, clinical guidelines recommend that health professionals decide upon the form of care by considering the expected course of the disease [18,60]. By deriving inferences of expected short-term disease developments from previously recorded symptom levels with our proposed model, health professionals can obtain decision support. If symptoms are expected to be subject to a reversal, then a low intensity form of care (or even pain self-management) might be sufficient. If symptoms are expected to persist, then health professionals may be required to choose a more intensive care.

As another implication, modeling the expected disease progression can also keep patients informed. This could allow them to adjust their disease self-management according to the most likely course of the disease [29]. For patients with low back pain, they can, for instance, perform specific exercises for pain self-management [25]. By inferring the expected course, patients can further adapt their daily actives such as work or travel accordingly. For diseases with substantial intrapersonal variability, it is often difficult to obtain accurate predictions.

The proposed model for decision support fulfills a key requirement from clinical settings that, due to involving high-stakes decisions, demand interpretability [54] to build trust [28]. First, the proposed model provides coefficients that allow for inference of a specific progression pattern (e.g., a tendency to persist or mean reverse) and corresponding prediction intervals. Second, it is based on a first-order autoregressive process, which considers a linear dependency between a previously recorded and expected symptomatic status. Third, through its cross-sectional fit, the model retains a parsimonious configuration, even across large patient populations. In the presented case, for example, our cross-sectional model yields 3 parameters in comparison to at least 2727 coefficients of patient-individual models but, despite that, with similar explanatory power. Altogether, our work fills a gap in prior literature (e.g., [26,58]) that has been largely focused on black-box approaches for modeling disease progression.

6.2. Implications for research

This work proposes the use of cross-sectional time series models in the domain of disease management. For this reason, the existing cross-sectional ARMA model was adapted in a way that extends to time series consisting of non-contiguous segments. From a research perspective, the cross-sectional time series models provide significant benefits: On the one hand, disease dynamics are inferred from a complete cohort of patients and should thus better reflect the different patients after deployment. On the other hand, predictions can be made for unseen patients (i.e., in so-called “offline” settings). To this end, the estimated coefficients from the rest of the patient cohort can be used. Consequently, there is no direct need to collect a sufficiently long time series for patients before offering predictions.

Our evaluations confirm that the proposed model captures a high amount of interpersonal variability of symptom progression ($R^2 = 53\%$ in both, explanatory and predictive setting). In comparison, many health professionals still adhere to the average level of a symptom as predictor [46]. As shown above, our model is superior to such mean-

Table 7
The above results extend the cross-sectional ARMA(1,1) model with an additional cross-sectional GARCH process in order to incorporate volatility clusters. The table again performs model selection with up to 5 model terms and lists the 3 best-performing models ordered by the AICc (preferred values in bold font). Here, the GARCH (2,0) is selected, since it removes the serial correlation as identified by the Ljung-Box test, assuming that residuals are independently distributed.

Model specification		Information criteria		Ljung-box test			
				Residuals		Squared residuals	
η	ζ	AICc	BIC	Test statistic	*P*-value	Test statistic	*P*-value
2	0	**167.07**	**176.38**	**0.08**	**0.77**	**1.73**	**0.19**
1	0	167.14	**174.75**	**1.70**	**0.19**	**18.90**	**0.00*****
3	0	168.44	179.34	0.04	0.84	0.01	0.94

Significance levels: *** 0.001, ** 0.01, * 0.05

Table 8
Parameter estimates of the cross-sectional GARCH(2,0) model and corresponding standard errors.

Parameter	Estimate
Intercept (ϕ)	0.824***
	(0.003)
First-order ARCH term (ψ_1)	0.349***
	(0.008)
Second-order ARCH term (ψ_2)	0.308***
	(0.006)

Stated: coefficients with standard errors in parentheses; significance levels are *** 0.001, ** 0.01, * 0.05.

models by a large extent. Additionally, by estimating our proposed model to data from specific cohorts, one has a simple possibility to personalize inferences to the disease dynamics of individual subgroups.

Our work proposes the utilization of autoregressive models for modeling expected disease trajectories. While autoregressive models have a long tradition of decision support on trajectories with high intrafactor variability, such as financial curves (e.g., [14,51]), they are less frequently applied in healthcare, where decisions are usually based on evidence from multiple subjects. Our cross-sectional autoregressive model thus fills a gap of suitable modeling approaches for healthcare analytics between linear regression and mean models on the one hand, and machine learning approaches on the other. Our model captures autoregression and can therefore make short-term predictions from disease trajectories with a high amount of intrapersonal variability, contrasting linear regression and mean models that capture variability only as variance. In opposition to more complex statistical models, such as Gaussian processes or neural networks, our cross-sectional ARMA model maintains interpretability.

Medical research benefits from our model as it helps in accumulating knowledge on the epidemiology of different diseases. More specifically, medical research can utilize our model to compare disease dynamics across different subgroups (e.g., to what extent disease dynamics are similar for females vs. males or patients with different characteristics, such as their mental wellbeing, degree of obesity, or disability). This might help in identifying subgroups where the disease dynamics differ from the overall population. Owing to the autoregressive structure, differences are quantified along dimensions that allow for straightforward interpretability (e.g., short-term reversals).

Prior research has advocated the use of longitudinal monitoring in disease management [5]. This can be achieved, for instance, via smart devices (e.g., smartwatches, wearable fitness trackers) or tailored apps for data collection. While tools for monitoring have become prevalent, the potential for more effective care has been largely untapped. One reason is that better models for decision support need to be developed [57]. In this regard, data-driven models have been primarily developed for applications such as risk scoring and the timing or dosage of medication, but they rarely guide the decision-making of health professionals in choosing a treatment plan. Our work provides a starting point for the latter: disease management from longitudinal health monitoring.

6.3. Limitations and generalizability

As with every modeling approach, ours is subject to assumptions. First, our model assumes the time series to be recorded in equidistant time steps. This is oftentimes the case when longitudinal monitoring (e.g., via smartwatches or the phone-based recording as in our work) is used. Longitudinal monitoring might yield missing values, yet this is successfully accommodated in our model specification. Second, our default model specification was intentionally set to a cross-sectional ARMA(1,1) model. While this might limit flexibility that could have been achieved from a higher-order lag structure, it achieves the desired level of interpretability. Notably, the information criteria in the above analysis point out that the information higher-order lag structures are on par with our proposed ARMA(1,1) model. Third, the input to our model is given by univariate patient data. Thereby, we follow extensive research in the area of decision support that is based on univariate modeling (e.g., [2,6,12,14]). In our work, the univariate specification was intentionally chosen to maintain interpretability. If desired, future research can extend to multivariate input.

The proposed model was evaluated based on low back pain. Nevertheless, its model specification should allow for broad applicability to other diseases that are long-lasting or even chronic. Examples of long-lasting conditions include headache disorders (where again, the input could be given by pain) or depression (where the severity over time is monitored). Meanwhile, chronic diseases include asthma, arthritis, or epilepsy. Owing to their chronic nature, disease management is again concerned with stabilizing the course of symptoms [17], so that stationarity is warranted.

As a first step, this work has laid the foundations for a modeling approach that could achieve relevance in practice. Our proposed model has been evaluated with one large patient sample, following common guidelines for predictive modeling [33]. Subsequently, future work will need to seek further validation with new data, and the model has to undergo impact testing as well as implementation evaluation.

7. Conclusion

For conditions that are long-lasting, clinical guidelines recommend that disease management adapts the intrapersonal variability in disease progression. Specifically, health professionals should consider the expected course of a disease in order to identify an effective form of care. This work has the potential to provide decision support by modeling the expected course of diseases in a cohort-wide setting. For this purpose, a cross-sectional, non-contiguous ARMA model is developed. This model differs from traditional time series models, as its parameters are estimated based on the complete patient cohort and thus generalize over multiple time series (rather than only one time series). Furthermore, our model adheres to requirements in practice, meaning it is based on a

parsimonious, theory-informed specification in order to ensure interpretability. Our proposed model is evaluated with longitudinal health data from a 52-week study involving 928 patients with low back pain. Health professionals can choose a low intensity form of care for patients where the pain level is low and expected to stabilize, whereas a more intensive treatment plan is required for patients with a high pain level and where persistence is expected. Thereby, our work facilitates disease management that is personalized to the intrapersonal disease dynamics of patients.

Author statement

Michael Mueller-Peltzer: Conceptualization, Methodology, Software, Formal analysis, Writing - Original Draft.

Stefan Feuerriegel: Supervision, Conceptualization, Methodology, Writing - Review & Editing.

Werner Vach: Investigation, Writing - Review & Editing.

Anne Molgaard Nielsen: Resources, Investigation, Writing - Review & Editing.

Alice Kongsted: Resources, Investigation, Writing - Review & Editing.

Dirk Neumann: Supervision, Writing - Review & Editing.

Acknowledgments

We thank Tim Howells for his professional language editing services. Stefan Feuerriegel gratefully acknowledges funding by the Swiss National Science Foundation (SNSF) as part of the SNSF Eccellenza grant 186932 “Data-driven health management”.

References

[1] G. Abraham, G.B. Byrnes, C.A. Bain, Short-term forecasting of emergency inpatient flow, IEEE Transactions on Information Technology in Biomedicine: A Publication of the IEEE Engineering in Medicine and Biology Society 13 (2009) 380–388.
[2] I.R. Adeyemi, S.A. Razak, M. Salleh, H.S. Venter, Observing consistency in online communication patterns for user re-identification, PLoS One 11 (2016) e0166930.
[3] S. Adeyemi, E. Demir, T. Chaussalet, Towards an evidence-based decision making healthcare system management: modelling patient pathways to improve clinical outcomes, Decis. Support. Syst. 55 (2013) 117–125.
[4] A. Allam, M. Nagy, G. Thoma, M. Krauthammer, Neural networks versus logistic regression for 30 days all-cause readmission prediction, Sci. Rep. 9 (2019) 9277.
[5] C.L. Anderson, R. Agarwal, The digitization of healthcare: boundary risks, emotion, and consumer willingness to disclose personal health information, Inf. Syst. Res. 22 (2011) 469–490.
[6] A.-L. Barabasi, The origin of bursts and heavy tails in human dynamics, Nature 435 (2005) 207–211.
[7] I. Bardhan, J.-H. Oh, Z. Zheng, K. Kirksey, Predictive analytics for readmission of patients with congestive heart failure, Inf. Syst. Res. 26 (2015) 19–39.
[8] D. Bertsimas, N. Kallus, A.M. Weinstein, Y.D. Zhuo, Personalized diabetes management using electronic medical records, Diabetes Care 40 (2017) 210–217.
[9] D. Bertsimas, A. O'Hair, S. Relyea, J. Silberholz, An analytics approach to designing combination chemotherapy regimens for cancer, Manag. Sci. 62 (2016) 1511–1531.
[10] T. Bollerslev, Generalized autoregressive conditional heteroskedasticity, J. Econ. 31 (1986) 307–327.
[11] P.J. Brockwell, R.A. Davis, Introduction to Time Series and Forecasting, 3rd ed., Springer, Cham, 2016.
[12] A. Carol, Market Risk Analysis, Pricing, Hedging and Trading Financial Instruments, John Wiley & Sons, Inc, 2008.
[13] CDC, Chronic disease overview, https://www.cdc.gov/chronicdisease/overview/index.htm, (2017).
[14] N.H. Chan, Time Series. Applications to Finance with R and S-Plus, 2nd ed., John Wiley & Sons, Inc, 2011.
[15] M.C. Chou, M. Parlar, Y. Zhou, Optimal timing to initiate medical treatment for a disease evolving as a semi-markov process, J. Optim. Theory Appl. 175 (2017) 194–217.
[16] J.H. Cochrane, Time Series for Macroeconomics and Finance, Manuscript University of Chicago, 2005, pp. 1–136.
[17] J. Corbin, A. Strauss, A nursing model for chronic illness management based upon the trajectory framework, Sch. Inq. Nurs. Pract. 5 (1991) 155–174.
[18] P. Croft, D.G. Altman, J.J. Deeks, K.M. Dunn, A.D. Hay, H. Hemingway, L. LeResche, G. Peat, et al., The science of clinical practice: disease diagnosis or patient prognosis? Evidence about what is likely to happen should shape clinical practice, BMC Med. 13 (2015) 20.
[19] A. Dag, A. Oztekin, A. Yucel, S. Bulur, F.M. Megahed, Predicting heart transplantation outcomes through data analytics, Decis. Support. Syst. 94 (2017) 42–52.
[20] K.K. Dobbin, R.M. Simon, Optimally splitting cases for training and testing high dimensional classifiers, BMC Med. Genet. 31 (2011) 1–8.
[21] A.S. Downie, M.J. Hancock, M. Rzewuska, C.M. Williams, C.-W.C. Lin, C.G. Maher, Trajectories of acute low back pain: a latent class growth analysis, Pain 157 (2016) 225–234.
[22] J.C. Driscoll, A.C. Kraay, Consistent covariance matrix estimation with spatially-dependent panel data, Rev. Econ. Stat. 80 (1998) 549–560.
[23] R.S. Epstein, L.M. Sherwood, From outcomes research to disease management: a guide for the perplexed, Ann. Intern. Med. 124 (1996) 832–837.
[24] R. Fisher, A. Smailagic, R. Simmons, K. Mizobe, Using latent variable autoregression to monitor the health of individuals with congestive heart failure, 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), 2016, pp. 1016–1019.
[25] N.E. Foster, J.R. Anema, D. Cherkin, R. Chou, S.P. Cohen, D.P. Gross, P.H. Ferreira, et al., Prevention and treatment of low back pain: evidence, challenges, and promising directions, Lancet 391 (2018) 2368–2383.
[26] J. Futoma, M. Sendak, B. Cameron, K. Heller, Predicting disease progression with a model for multivariate longitudinal clinical data, Machine Learning for Healthcare Conference, 2016, pp. 42–54.
[27] M. Ghassemi, M.A.F. Pimentel, T. Naumann, T. Brennan, D.A. Clifton, P. Szolovits, M. Feng, A multivariate timeseries modeling approach to severity of illness assessment and forecasting in ICU with sparse, heterogeneous clinical data, Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015, pp. 446–453.
[28] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM Comput. Surv. 51 (2019) 93.
[29] A. Gupta, R. Sharda, Improving the science of healthcare delivery and informatics using modeling approaches, Decis. Support. Syst. 55 (2013) 423–427.
[30] M.A. Hamburg, F.S. Collins, The path to personalized medicine, N. Engl. J. Med. 363 (2010) 301–304.
[31] J. Hartvigsen, M.J. Hancock, A. Kongsted, Q. Louw, M.L. Ferreira, S. Genevay, D. Hoy, et al., What low back pain is and why we need to pay attention, Lancet 391 (2018) 2356–2367.
[32] A.C. Harvey, Time Series Models, Harvester Wheatsheaf, New York, 1993.
[33] T. Hastie, R. Tibshirani, J.H. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed., Springer Series in Statistics, Springer, New York, 2009.
[34] H. Hemingway, P. Croft, P. Perel, J.A. Hayden, K. Abrams, A. Timmis, A. Briggs, et al., Prognosis research strategy (PROGRESS) 1: a framework for researching clinical outcomes, BMJ (Clinical Research Ed.) 346 (2013) e5595.
[35] J. Hlouskova, M. Wagner, The performance of panel unit root and stationarity tests: results from a large scale simulation study, Econ. Rev. 25 (2006) 85–116.
[36] T.M.H. Hope, M.L. Seghier, A.P. Leff, C.J. Price, Predicting outcome and recovery after stroke with lesions extracted from MRI images, NeuroImage: Clinical 2 (2013) 424–433.
[37] Z. HU, Y. Bao, R. Chiong, T. Xiong, Mid-term interval load forecasting using multi-output support vector regression with a memetic algorithm for feature selection, Energy 84 (2015) 419–431.
[38] C.M. Hurvich, C.-L. Tsai, A corrected Akaike information criterion for vector autoregressive model selection, J. Time Ser. Anal. 14 (1993) 271–279.
[39] S.L. James, et al., Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the global burden of disease study 2017, Lancet 392 (2018) 1789–1858.
[40] A. Kongsted, P. Kent, I. Axen, A.S. Downie, K.M. Dunn, What have we learned from ten years of trajectory research in low back pain? BMC Musculoskelet. Disord. 17 (2016) 220.
[41] A. Kongsted, P. Kent, L. Hestbaek, W. Vach, Patients with low back pain had distinct clinical course patterns that were typically neither complete recovery nor constant pain: a latent class analysis of longitudinal data, Spine J. 15 (2015) 885–894.
[42] R.J. Larse, M. Kasimatis, Day-to-day physical symptoms: individual differences in the occurrence, duration, and emotional concomitants of minor daily illnesses, J. Pers. 59 (1991) 387–423.
[43] M.J. Lebo, C. Weber, An effective approach to the repeated cross-sectional design, Am. J. Polit. Sci. 59 (2015) 242–258.
[44] A. Levin, C.-F. Lin, C.-S.J. Chu, Unit root tests in panel data: asymptotic and finite-sample properties, J. Econ. 108 (2002) 1–24.
[45] Y.-K. Lin, H. Chen, R.A. Brown, S.-H. Li, H.-J. Yang, Time-to-event predictive modeling for chronic conditions using electronic health records, IEEE Intell. Syst. 29 (2014) 14–20.
[46] C.J. Mun, H.W. Suk, M.C. Davis, P. Karoly, P. Finan, H. Tennen, M.P. Jensen, Investigating intraindividual pain variability: methods, applications, issues, and directions, Pain 160 (2019) 2415–2429.
[47] D.M. Negoescu, K. Bimpikis, M.L. Brandeau, D.A. Iancu, Dynamic learning of patient response types: an application to treating chronic diseases, Manag. Sci. 64 (2018) 3469–3970.
[48] J. Ni, X. Jin, Decision support systems for effective maintenance operations, CIRP Ann. 61 (2012) 411–414.
[49] A.M. Nielsen, A. Binding, C. Ahlbrandt-Rains, M. Boeker, S. Feuerriegel, W. Vach, Exploring conceptual preprocessing for developing prognostic models: a case study in low back pain patients, J. Clin. Epidemiol. (2020) (forthcoming).
[50] A.M. Nielsen, W. Vach, P. Kent, L. Hestbaek, A. Kongsted, Using existing questionnaires in latent class analysis: should we use summary scores or single items as input? A methodological study using a cohort of patients with low back pain, Clinical Epidemiology 8 (2016) 73–89.
[51] J. Nowotarski, R. Weron, Computing electricity spot price prediction intervals using

quantile regression and forecast averaging, Comput. Stat. 30 (2015) 791–803.
[52] S. Piri, D. Delen, T. Liu, H.M. Zolbanin, A data analytics approach to building a clinical decision support system for diabetic retinopathy: developing and deploying a model ensemble, Decis. Support. Syst. 101 (2017) 12–27.
[53] H. Quan, D. Srinivasan, A. Khosravi, Uncertainty handling using neural network-based prediction intervals for electrical load forecasting, Energy 73 (2014) 916–925.
[54] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature Machine Intelligence 1 (2019) 206–215.
[55] S. Safikhani, K.S. Gries, J.J. Trudeau, D. Reasner, K. Rüdell, S.J. Coons, E.N. Bush, J. Hanlon, L. Abraham, M. Vernon, Response scale selection in adult pain measures: results from a literature review, Journal of Patient-Reported Outcomes 2 (2017) 40.
[56] L.W. Sayrs, Pooled Time Series Analysis, Sage Publications, Inc, 1989.
[57] S. Scholtes, C. Terwiesch, Empirical research in healthcare operations: past research, present understanding, and future opportunities, Manuf. Serv. Oper. Manag. 22 (1) (2020) 73–83.
[58] P. Schulam, S. Saria, A framework for individualizing predictions of disease trajectories by exploiting multi-resolution structure, Advances in Neural Information Processing 2015 (2015) 748–756.
[59] R. Shibl, M. Lawley, J. Debuse, Factors influencing decision support system acceptance, Decis. Support. Syst. 54 (2013) 953–961.
[60] E.W. Steyerberg, K.G.M. Moons, D.A. van der Windt, J.A. Hayden, P. Perel, S. Schroter, R.D. Riley, et al., Prognosis research strategy (PROGRESS) 3: prognostic model research, PLoS Med. 10 (2013) e1001381.
[61] L.A. Thombs, W.R. Schucany, Bootstrap prediction intervals for autoregression, J. Am. Stat. Assoc. 85 (1990) 486–492.
[62] H. Wang, N. Feping, H. Huang, Y. Jingwen, K. Sungeun, S.L. Risacher, A.J. Saykin, L. Shen, High-order multi-task feature learning to identify longitudinal phenotypic markers for Alzheimer's disease progression prediction, Advances in Neural Information Processing 2012, 2012, pp. 1277–1285.
[63] C.S. Wong, W.K. Li, A note on the corrected Akaike information criterion for threshold autoregressive models, J. Time Ser. Anal. 19 (1998) 113–124.
[64] H.M. Zolbanin, D. Delen, Processing electronic medical records to improve predictive analytics outcomes for hospital readmissions, Decis. Support. Syst. 112 (2018) 98–110.
[65] H.M. Zolbanin, D. Delen, A. Hassan Zadeh, Predicting overall survivability in comorbidity of cancers: a data mining approach, Decis. Support. Syst. 74 (2015) 150–161.

Michael Mueller-Peltzer is a Ph.D. student at the Chair of Information Systems of the University of Freiburg, Germany. His research focuses on decision support for data-driven health management. Previously, he worked as a senior consultant in the field of healthcare for the global management consulting company McKinsey & Company. He obtained a diploma in medical engineering from Technical University Munich (TUM) in 2014.

Stefan Feuerriegel is an assistant professor for management information systems at ETH Zurich. His research focuses in particular on better decision support for data-driven health management. Previously, he obtained his Ph.D. from the University of Freiburg. He has co-authored research publications in the European Journal of Operational Research, the European Journal of Information Systems, the Journal of Information Technology and Decision Support Systems.

Werner Vach is a senior researcher at the Department of Orthopaedics and Traumatology, University Hospital Basel. His research focuses on the interplay between medical statistics and medical research. He has co-authored more than 50 publications in methodological journals and more than 200 publications in various fields of clinical research.

Anne Molgaard Nielsen is a teaching assistant professor in clinical biomechanics at the University of Southern Denmark, from where she also obtained her Ph.D. Before that she was a practicing chiropractor for 4 years in a private clinic and 1 year in a specialized hospital clinic. Her research focuses on musculoskeletal health with a special interest in low back pain. She has co-authored research publications in the BMC Musculoskeletal Disorders, Journal of Physiotherapy, Clinical Epidemiology and the Scandinavian Journal of Medicine & Science in Sports.

Alice Kongsted is a professor at the Department of Sports Science and Clinical Biomechanics at the University of Southern Denmark and a senior researcher at the Nordic Institute of Chiropractic and Clinical Biomechanics. Her research concerns spinal pain with a focus on primary care management. She is an Associate Editor of Chiropractic & Manual Therapies. In 2018, she was part of the Lancet Low Back Pain Series Working Group.

Dirk Neumann is a professor at the Chair of Information Systems of the University of Freiburg, Germany. His research topics include Business Analytics, Text Mining and Cloud Computing. He studied information systems in Giessen (Diploma), Economics in Milwaukee, WI, USA (Master) and received a PhD from Karlsruhe Institute of Technology (KIT) in 2004. He has (co-)authored many research publications at European Journal of Operational Research, ACM Transactions on Internet Technology, Journal of Management Information Systems or Decision Support Systems.