Predictive Modeling of Therapy Decisions in Metastatic Breast Cancer with Recurrent Neural Network Encoder and Multinomial Hierarchical Regression Decoder

Yinchong Yang, Volker Tresp
Ludwig-Maximilians-Universität München
Siemens AG, Corporate Technology, Munich
{yinchong.yang, volker.tresp}@siemens.com

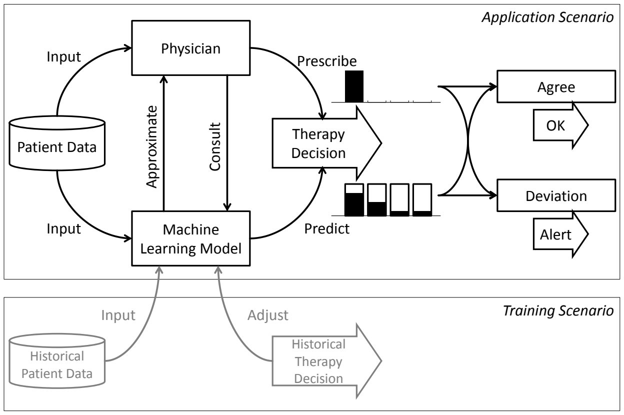
Peter A. Fasching
Department of Gynecology and Obstetrics,
University Hospital Erlangen
peter.fasching@uk-erlangen.de

***Abstract*—The increasing availability of novel health-related data sources —e.g., from molecular analysis, health Apps and electronic health records— might eventually overwhelm the physician, and the community is investigating analytics approaches that might be useful to support clinical decisions. In particular, the success of the latest developments in Deep Learning has demonstrated that machine learning models are capable of handling —and actually profiting from— high dimensional and possibly sequential data. In this work, we propose an encoder-decoder network approach to model the physician's therapy decisions. Our approach also provides physicians with a list of similar historical patient cases to support the recommended decisions. By using a combination of a Recurrent Neural Network Encoder and a Multinomial Hierarchical Regression Decoder, we specifically tackle two common challenges in modeling clinical data: First, the issue of handling episodic data of variable lengths and, second, the need to represent hierarchical decision procedures. We conduct experiments on a large real-world dataset collected from thousands of metastatic breast cancer patients and show that our model outperforms more traditional approaches.**

1. Introduction

With the introduction of the Electronic Health Records (EHR), a large amount of digital information has become available in clinics. This is expected to encourage more personal and precise healthcare services and improve patients experience [1, 2]. On the other hand, it also requires the physicians to consult a large variety and volume of data in order to perform diagnosis and treatment decisions, such as the patients' background information, medical images, genetic profiles and the patients' entire medical history. The decision making process, therefore, could become increasingly complex in connection with the growing amounts of information collected on each patient. Machine learning based Clinical Decision Support (CDS) systems could provide a solution to such data challenges [3, 4, 5]. These systems are able to actually profit from the large amount of data in high dimensional space. For instance, the latest success of Deep Learning models lies in their ability to generate more abstract and informative latent features from the high dimensional raw features, which turns out to largely facilitate predictive modeling.

Figure 1. The concept of a machine learning based CDS system.



There are multiple ways that a machine learning model may impact the decision process of a physician, for instance, by predicting the possible outcome of each decision. [5] provides physicians with endpoint predictions of patients with kidney failure. Based on the predicted probabilities of kidney rejection, kidney loss and death within the next 6 and 12 months, the physician is more informed to select the correct medication. This class of approaches, however, might be limited when i) not yet enough endpoints are labeled in the training data and ii) confounder effect is presumed in the data situation [6]. Therefore, in this work we explore another approach to machine learning base decision support by directly predicting the physicians' decisions. More specifically, a machine learning model would calculate the probability of each decision conditioned on the patient information. From the viewpoint of the physicians, these probabilities can be interpreted as recommendation scores. We illustrate this conceptual framework in Fig. 1. When properly trained, the machine learning model can be expected to generate recommendations which –to a certain extend– agree with the prescriptions actually prescribed by the committee of physician in the study. In cases where the physician faces a great number of possible decisions, the recommendations would narrow down the size of prescription candidates. On the other hand, the machine learning model would also implicitly detect anomalous prescriptions, by checking whether the actual prescriptions are among the top-n ranked recommendations made by the machine learning model. Such a system relies on the predictive power of the machine learning model, which can be trained using historical data. During training, the model attempts to predict historical decisions based on the corresponding patient data and the actually documented decisions can adjust the model so that it can improve its predictions throughout the training epochs.

Our study is based on a large and up-to-date data set consisting of almost three thousand metastatic breast cancer patients in Germany. This dataset includes the patients' background information, the primary tumor record, etc., as well as the development of the cancer such as local recurrence and metastasis. Included in the dataset are also all the prescribed treatments each patient obtained throughout time. Since the physicians make their therapy decisions —often at a tumor board— after studying all available patient information, we assume that a machine learning model can also be trained to map the patient features to the therapy decisions.

There are two major challenges in the data situation. Firstly, the patients in the dataset do not share a time axis and do not visit clinics regularly. On some patients, no more data was ever recorded after a surgery; while others revisited the clinics repeatedly for years due to local recurrences and metastasis. Consequently, the patients have a medical history of variable length, making it challenging to construct a common input feature space for all patients. Secondly, the therapy decisions that we attempt to model are of a hierarchical structure. For instance, the physician first has to decide for a radiotherapy before further specifying whether it should be a curative or a palliative one, and whether it should be a Brachytherapy or a percutaneous type.

To address these two issues we propose a neural network architecture that instantiates the Encoder-Decoder Framework by [7]. Specifically, we encode the patients' medical histories of *variable lengths* into one *fixed-size* representation vector using Recurrent Neural Networks (RNN), and deploy on top of that a hierarchical regression model, which functions as a decoder that predicts the therapy decisions. We conduct experiments on the dataset with multiple choices of encoders and decoders, as well as different hyper-parameter settings, and identify their contribution to the modeling quality. Furthermore, we show that with our model architecture, one could also provide physicians with a list of similar historical patient cases to support our prediction, making it more realistic to deploy such decision support system in clinics.

The rest of the paper is organized as follows. In Section 2 we discuss multiple related works that inspired the design of our model. In Section 3 we describe our data situation, including the study background and data processing approaches. In Section 4 we first briefly introduce two specific RNN models that serve as our encoder network, and then propose a hierarchical prediction model as our decoder. In Section 5 we present our experimental results and Section 6 wraps up our present work and give a outlook for future directions.

2. Related Work

Handling sequential EHR data. Due to the sequential nature of EHR data, there have recently been multiple promising works studying clinical events as sequential data. Many of them are inspired by works in natural language modeling, since sentences can be easily modeled as sequence of signals. [4] adjusted a language model based on the sliding window technique in [8], taking into account a fixed number of events in the past. This model was extended in [5] by replacing the sliding window with RNNs, which improved the predictions for prescriptions decision and endpoints. [9] also applied LSTM-RNN to perform diagnosis prediction based on sequential input. And a related approach with RNNs can also be found in [3] to predict diagnosis and medication prescriptions at the same time throughout time. Such RNN application was further augmented with neural attention mechanism in [10], which did not only show promising results but also improved the interpretability of the model.

RNNs for sequence classification/regression. The RNN models in these works were implemented in a many-to-many fashion. That is to say, at each time step the RNN is supposed to generate a prediction as output. The reason is that in their data all patients share the same aligned time axis and regularly visit the clinics. In our work, on the other hand, there are neither regular visits nor shared time axis. To this end we implemented many-to-one RNN models that consume a sequential input and generates only one output. This setting can be found in a variety of sequence classification/regression tasks. [11] used such RNN architectures to classify spoken words and handwriting as sequences. RNNs have been also applied to classify the sentiment of a sentence such as in the IMDB reviews dataset [12]. The applications of the RNNs in the many-to-one fashion can also be seen as the encoding of a sequence of variable length into one fixed-size representation [7], which is then decoded to perform prediction as decoding.

Hierarchical classification/regression model. Rather than a simple classification task where all classes are on the same level, the therapy decisions turn out to be more complicated. For instance, the decision of a Brachytherapy is only observed when the physician decides to prescribe a radiotherapy in the first place. In order to model such a decision procedure as realistic as possible, we extend a hierarchical response model in [13] and deploy it as decoder on top of RNNs. [14] also proposed a quite similar architecture to factorize a large softmax layer into a hierarchy. The purpose was to accelerate the calculation of the softmax, which in natural language processing often has the size of the entire vocabulary.

3. Metastatic Breast Cancer Data

In this section we first briefly introduce the classical breast cancer therapies and then give an overview of our data situation.

3.1. Metastatic Breast Cancer Treatments

Breast cancer is the one of the most frequent malignant cancers in the Western world. In Germany, for instance, approximately 70,000 women suffer from breast cancer each year with around 30% mortality rate [15, 16]. In many of these cases, it is the metastasis of the cancer cells to vital organs that actually causes the patient's death. There are three classes of classical treatments of metastatic breast cancer: radiotherapy, systemic therapy and surgery. Typically, as soon as a patient is diagnosed with breast cancer, a surgery to remove the primary tumor would be the first option. In order to prevent local recurrence and metastasis, the patient would receive radiotherapies and/or systemic therapies after the surgery. If, however, local recurrences and/or metastasis are later diagnosed, the patient might undergo a further surgery, succeeded by radiotherapies and/or systemic therapies. This process can be repeated till either i) no more recurrence or metastasis can be identified or ii) the metastasis is observed in vital organs and surgery is no longer an option. In the latter case, radiotherapies and/or systemic therapies would become the major treatments. Latest discoveries in genetics have brought about novel systemic therapies that exploit specific biological characteristics of the cancer cell. Since these special characteristics are mostly not present in healthy cells, these targeted therapies have proven to be more efficient with less severe adverse effect.

3.2. Data Description and Processing

The majority of the dataset was provided by the PRAEGNANT study network [17], which has been recruiting patients of metastatic breast cancer since 2014. The original data are warehoused in the secuTrial® database. After exporting and pre-processing, we could extract information on 2,869 valid patients.

There are two classes of patient information that are potentially relevant for modeling the therapy decisions: First the *static* information includes 1) basic patient properties, 2) information on the primary tumor and 3) information on the history of metastasis before entering the study. In total we observe 26 features of binary, categorical or real types. We performed dummy-coding on the former both cases and could extract for each patient i a static feature vector denoted with $\boldsymbol{m}_i \in \mathbb{R}^{118}$. We summarize the features in Tab. 1.

The *sequential* information includes data on 4) local recurrences, 5) metastasis 6) clinical visits 7) radiotherapies, 8) systemic therapies and 9) surgeries. These are time-stamped clinical events observed on each patient throughout time, and at each time step there can be more than one type of events recorded. All these sequential features are of binary or categorical nature and are also dummy-coded, yielding for patient i at time step t a feature vector $\boldsymbol{x}_i^{[t]} \in \{0,1\}^{189}$. We denote the whole sequence of events for this patient i up to time T_i using a set of $\{\boldsymbol{x}_i^{[t]}\}_{t=1}^{T_i}$. We summarize the sequential features in Tab. 2.

Table 1. OVERVIEW OF ALL STATIC FEATURES.

Static features	Feature names and dimensions	
1) Basic	Age	1
	Height	1
	HRT (Hormone Replacement Therapy)	5
	parity	9
	Mother BC	3
	Sister BC	6
	Menstruation	1
2) Primary Tumor	Type	3
	Total eval. of the malignancy	8
	Total eval. of axilla	4
	TAST eval. of the malignancy	8
	TAST eval. of axilla	4
	Mammography eval. of the malignancy	8
	Mammography eval. of axilla	4
	Ultrasound eval. of the malignancy	8
	Ultrasound eval. of axilla	4
	MRI eval. of the malignancy	8
	MRI eval. of axilla	8
	Metastasis staging	4
	Ever neoadjuvant therapy	4
	Ever surgery	4
3) History of metastasis	Lungs	1
	Liver	1
	Bones	1
	Brain	1
	Other	10
Total	26	118

Since we attempt to model the therapy decisions concerning radiotherapies (item 7), systemic therapies (item 8)[1] and surgeries (item 9), we extract from the medical history of each patient all possible sub-sequences where the last event consists of one of the three therapies. Therefore in each of these sub-sequences, the last event serves as the target that the model is expected to predict based on all previous events and the static information. Obviously, instead of the entire vector $\boldsymbol{x}_i^{[t]}$ we only need the subset of the vector concerning the therapies and denote this with $\boldsymbol{y} \in \{0,1\}^{39}$. Finally the training/test samples are constructed as

$$\{\boldsymbol{m}_i, \{\boldsymbol{x}_i^{[t]}\}_{t=1}^{t_i^*-1}\} \to \boldsymbol{y}_i^{[t_i^*]} \subseteq \boldsymbol{x}_i^{[t_i^*]}, \tag{1}$$

for each possible time step t_i^* where one of the therapies is observed. We illustrate this approach of data processing in Fig. 2.

From the 2,869 patients we could extract in total 16,314 sequences (i.e. 5.7 sequence per patient on average). The length of the sequence before a therapy prescription varies from 0 to 35 and is on average 4.1.

Every time a physician is supposed to prescribe a treatment, she/he is first supposed to choose one of the three *therapy categories* of radiotherapy, systemic therapy and surgery. For each chosen therapy category the physician

[1] Except the fourth feature ”Reason of termination”, which we do not deem predictable but would serve well as input feature.

Table 2. OVERVIEW OF ALL SEQUENTIAL FEATURES, THEREOF 7), 8) AND 9) ARE THERAPIES THAT WE ATTEMPT TO PREDICT.

Sequential features	Feature names and dimensions	
4) Local Recurrences	Location	4
	Type	3
5) Metastasis Evaluation	Total	6
	Lungs	9
	Liver	9
	Bones	9
	Brain	9
	Lymph	9
	Skin	9
	Ovum	9
	Soft tissue	8
	Kidney	8
	Pleural cavity	8
	Thorax	8
	Muscle	8
	Periosteum	8
	Other	8
6) Visits	Therapy situation	12
	ECOG Life status	6
7) Radiation	Type	3
	Intention	3
8) Systemic	Type	6
	Intention	13
	Ref. to an surgery	4
	Reason of termination	6
9) Surgery	Type	10
Total	26	189

will then decide the *therapy features*. For radiotherapy there are two 3-dimensional multinomial distributed features: the radiotherapy intention being either curative, palliative or unknown; and the radiotherapy's type being either percutaneous, Brachytherapy or others. For systemic therapy there are three multinomial distributed features. The first one describes 6 types of systemic therapy such as antihormone therapy, chemotherapy, anti-HER2 therapy etc.; the second feature documents the therapy's intention, namely an argument based on the 13 different stagings of the cancer; the third four-dimensional feature records whether the therapy prescription is related to a surgery or is unknown. The last category is composed of 10 Bernoulli distributed variables that describe the surgery, such as breast conservation surgery, mastectomy, etc.. Detailed information of the feature *values* can be found in Tab. 3.

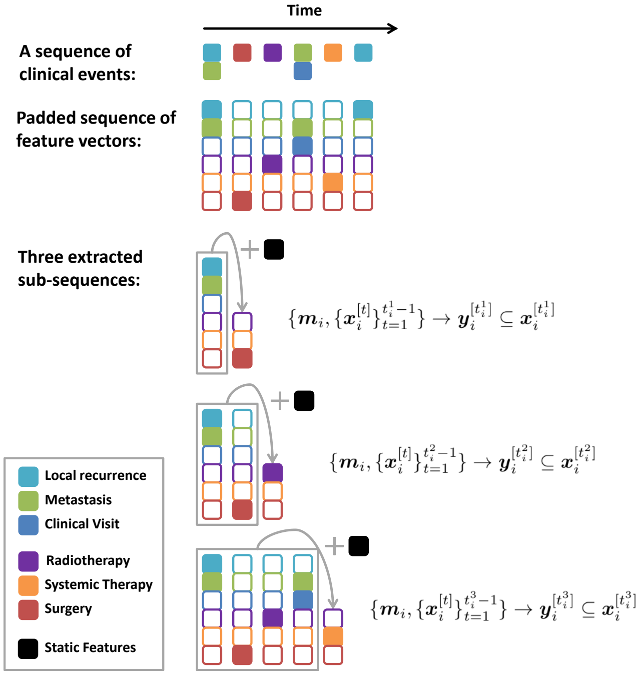
4. A Predictive Model of Therapy Decisions

In this section we provide an introduction to the two core ingredients of our proposed model: the many-to-one RNNs and a Multinomial Hierarchical Regression model. Eventually, both will be joined to form the complete predictive model.

4.1. Recurrent Neural Network as Encoder

Recurrent Neural Networks, especially the more advanced variants of Gated Recurrent Units (GRU) [18], presented in Eq. (2), and Long Short-Term Memory (LSTM) [19, 20] as in Eq. (3) have proven to be powerful in modeling multidimensional sequential data such as sensory and natural language data [21, 22].

Figure 2. Illustration of generating training and test sequences from the medical history of a patient. From a complete sequence of clinical events, we extract all possible sub-sequences that end with one or multiple therapies, in this case at t_i^1, t_i^2 and t_i^3. At each time step, if a specific event is not observed, its corresponding features are zero-padded, yielding a common feature space at each time step.


Time
A sequence of clinical events:
Padded sequence of feature vectors:
Three extracted sub-sequences:
$\{m_i, \{x_i^{[t]}\}_{t=1}^{t_i^1-1}\} \to y_i^{[t_i^1]} \subseteq x_i^{[t_i^1]}$
$\{m_i, \{x_i^{[t]}\}_{t=1}^{t_i^2-1}\} \to y_i^{[t_i^2]} \subseteq x_i^{[t_i^2]}$
$\{m_i, \{x_i^{[t]}\}_{t=1}^{t_i^3-1}\} \to y_i^{[t_i^3]} \subseteq x_i^{[t_i^3]}$
Local recurrence
Metastasis
Clinical Visit
Radiotherapy
Systemic Therapy
Surgery
Static Features

GRU:

$$\begin{aligned} \boldsymbol{r}^{[t]} &= \sigma(\boldsymbol{W}^r\boldsymbol{x}^{[t]} + \boldsymbol{U}^r\boldsymbol{h}^{[t-1]} + \boldsymbol{b}^r) \\ \boldsymbol{z}^{[t]} &= \sigma(\boldsymbol{W}^z\boldsymbol{x}^{[t]} + \boldsymbol{U}^z\boldsymbol{h}^{[t-1]} + \boldsymbol{b}^z) \\ \boldsymbol{d}^{[t]} &= \tanh(\boldsymbol{W}^d\boldsymbol{x}^{[t]} + \boldsymbol{U}^d(\boldsymbol{r}^{[t]} \circ \boldsymbol{h}^{[t-1]})) \\ \boldsymbol{h}^{[t]} &= (1 - \boldsymbol{z}^{[t]}) \circ \boldsymbol{h}^{[t-1]} + \boldsymbol{z}^{[t]} \circ \boldsymbol{d}^{[t]}, \end{aligned} \tag{2}$$

LSTM:

$$\begin{aligned} \boldsymbol{k}^{[t]} &= \sigma(\boldsymbol{W}^k\boldsymbol{x}^{[t]} + \boldsymbol{U}^k\boldsymbol{h}^{[t-1]} + \boldsymbol{b}^k) \\ \boldsymbol{f}^{[t]} &= \sigma(\boldsymbol{W}^f\boldsymbol{x}^{[t]} + \boldsymbol{U}^f\boldsymbol{h}^{[t-1]} + \boldsymbol{b}^f) \\ \boldsymbol{o}^{[t]} &= \sigma(\boldsymbol{W}^o\boldsymbol{x}^{[t]} + \boldsymbol{U}^o\boldsymbol{h}^{[t-1]} + \boldsymbol{b}^o) \\ \boldsymbol{g}^{[t]} &= \tanh(\boldsymbol{W}^g\boldsymbol{x}^{[t]} + \boldsymbol{U}^g\boldsymbol{h}^{[t-1]} + \boldsymbol{b}^g) \\ \boldsymbol{c}^{[t]} &= \boldsymbol{f}^{[t]} \circ \boldsymbol{c}^{[t-1]} + \boldsymbol{k}^{[t]} \circ \boldsymbol{g}^{[t]} \\ \boldsymbol{h}^{[t]} &= \boldsymbol{o}^{[t]} \circ \tanh(\boldsymbol{c}^{[t]}). \end{aligned} \tag{3}$$

Both models generate for each time stamped input $\boldsymbol{x}^{[t]}$ a hidden state $\boldsymbol{h}^{[t]}$ that depends on both the current input $\boldsymbol{x}^{[t]}$ and the last representation $\boldsymbol{h}^{[t-1]}$.

If one has a sequence of targets $\{\boldsymbol{y}^{[t]}\}_{t=1}^{T}$ with the same length as $\{\boldsymbol{x}^{[t]}\}_{t=1}^{T}$ (a many-to-many model) such as in [3, 5], one could build a prediction model on top of every hidden state: $\hat{\boldsymbol{y}}^{[t]} = \phi(\boldsymbol{h}^{[t]})$ $\forall t$. On the other hand, one could also

Table 3. THE MODELING TARGET: THERAPY CATEGORIES, THERAPY FEATURES AND FEATURE VALUES.

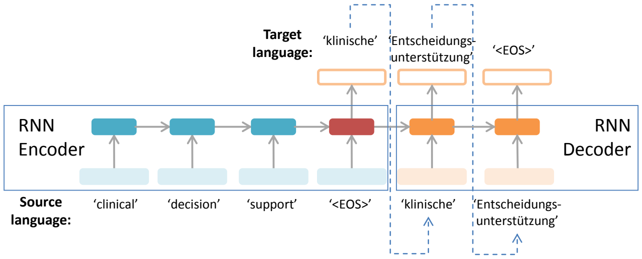
Therapy category	Therapy feature	Feature value
7) Radiation	Intention	Curative Palliative unknown
	Type	percutaneous radiation Brachytherapy Other radiotherapies
8) Systemic	Type	Anti-hormone therapy Chemotherapy Anti-HER2 therapy Other anti-body therapy Bone specific therapy Other therapies
	Intention	CM0/First treatment. CM0/Treatment of local recc. 1st line met 2nd line met 3rd line met 4th line met 5th line met 6th line met 7th line met 8th line met 9th line met not filled unknown
	Ref. to surgery	Neoadjuvant Adjuvant No surgery Unknown
9) Surgery	Type	Breast-Conserving Therapy Mastectomy Excision Trial Sampling Diagnostic Sampling Sentinel-Node-Biopsy Skin Sparing Mastectomy Port-Implantation Paracentesis Reconstruction

have a many-to-one model with only one target y for the whole input sequence of $\{\boldsymbol{x}^{[t]}\}_{t=1}^{T}$. In such case a prediction model consumes the last hidden state, which recurrently depends on all its predecessors, in form of $\hat{\boldsymbol{y}} = \phi(\boldsymbol{h}^{[T]})$.

Interestingly, [7] proposed a Encoder-Decoder-Framework for machine translation, which involves both of these variants. First a many-to-one RNN encodes a sentence of the source language into its last hidden state vector, which is interpreted as the representation for the entire sentence. The second one is a many-to-many RNN. It consumes the last hidden state of the encoder as its first hidden state and generates a sentence of the target language. We illustrate this model using a simple example in Fig. 3.

In their work the many-to-one RNN was proven capable of learning a fixed-size representation from the entire input sequence of variable length, which is an appealing characteristics for our data situation as well. In our data, each patient case has a medical history of variable length, and the number of clinical events observed before a therapy prescription varies between 0 and 35. With such an encoder

Figure 3. The Encoder-Decoder-Framework for Machine Translation by [7]. The encoder RNN outputs only its last hidden state when it sees the end-of-sentence symbol. At the same time, the decoder RNN consumes this hidden state as its initial one and generates the first word. The decoder keeps generating words till it generates the end-of-sentence symbol.



RNN we could extract from such sequential input a more abstract and compact vector representing the entire history of the patient up to a specific time step. For the sake of simplicity, we denote such a many-to-one RNN (either GRU or LSTM) using a function ω:

$$\boldsymbol{h}_i^{[t^*]} = \omega(\{\boldsymbol{x}_i^{[t]}\}_{t=1}^{t^*}), \tag{4}$$

where $\boldsymbol{h}_i^{[t^*]}$ is the last hidden state.

In order to also take into account the static features such as patient information and primary tumor, we follow [5] and concatenate the output of the RNN with the latent representation of the static features.

$$\boldsymbol{z}_i^{[t^*]} = (\boldsymbol{h}_i^{[t^*]}, \boldsymbol{q}_i) \text{ with } \boldsymbol{q}_i = \psi(\boldsymbol{H}^T \boldsymbol{m}_i), \tag{5}$$

where $\boldsymbol{H}$ is a usual trainable weight matrix and ψ denotes a non-linear activation function. Therefore, the vector $\boldsymbol{z}_i^{[t^*]}$ represents the static patient information as well as the medical history of patient i up to time step t^*. Such a vector functions as an abstract patient profile that represents all relevant clinical information in a latent vector space, where patients with similar background information and medical history could be encouraged to be placed in a specific neighborhood. This very characteristic of the latent vector space is key to the latest success of Deep Learning, in that it facilitates the classification and regression models built on top of it.

4.2. Hierarchical Response Model as Decoder

We attempt to model the therapies in a similar fashion as the physicians' prediction procedure. A physician first has to choose one therapy category, and then to specify for the chosen category its features. We propose a Multinomial Hierarchical Regression (MHR) to model this procedure.

In the first step we model the probability that each of the three therapy categories is chosen at time step t^* for patient i using a multinomial variable $C_i^{[t^*]}$ with a softmax activation:

$$\mathbb{P}(C_i^{[t^*]} = k \mid \boldsymbol{m}_i, \{\boldsymbol{x}_i^{[t]}\}_{t=1}^{t^*}) = \frac{\exp\left((\boldsymbol{z}_i^{[t^*]})^T \boldsymbol{\gamma}^k\right)}{\sum_{\forall k'} \exp\left((\boldsymbol{z}_i^{[t^*]})^T \boldsymbol{\gamma}^{k'}\right)}, \tag{6}$$

where $\boldsymbol{z}_i^{[t^*]}$, as defined in Eq. (5), is the latent representation for the patient up to this time step and $\boldsymbol{\gamma}^k$ serves as the category-specific parameter vector.

Then in the second step, given a specific therapy category k, we denote the number of therapy features in this category with L^k and model the l^k-th multinomial distributed feature variable F_{k,l^k}, whose conditional probability can be modeled with

$$\begin{aligned} &\mathbb{P}(F_{i,k,l^k}^{[t^*]} = r \mid C_i^{[t^*]} = k, \boldsymbol{m}_i, \{\boldsymbol{x}_i^{[t]}\}_{t=1}^{t^*}) = \\ =&\frac{\exp\left((\boldsymbol{z}_i^{[t^*]})^T \boldsymbol{\beta}_{k,l^k,r}\right)}{\sum_{\forall r'} \exp\left((\boldsymbol{z}_i^{[t^*]})^T \boldsymbol{\beta}_{k,l^k,r'}\right)}, \end{aligned} \tag{7}$$

if k=1 or k=2, i.e. in case of radiotherapy or systemic therapy where therapy features in each category are multiple multinomial distributed. Therefore one would need the softmax function to model the probabilities that the therapy feature takes one specific value r. We denote the parameter vector $\boldsymbol{\beta}_{k,l^k,r}$ with three levels of subscripts: k suggests the category of the therapy, l^k selects one specific multinomial feature from this category, and r denotes the r-th possible outcome of this feature. For instance, we would use $\boldsymbol{\beta}_{1,2,3}$ to denote the parameters corresponding to the hierarchy of `radiotherapy` / `type` / `other_radiotherapy`, implying that the type of the radiotherapy is of other kinds (3rd column, 6th row in Tab. 3).

If the therapy category suggests the surgery, i.e. k=3, whose features consist of L_k=10 Bernoulli variable, we would have instead of Eq. (7) the following formulation:

$$\begin{aligned} &\mathbb{P}(F_{i,k,l^k}^{[t^*]} = r \mid C_i^{[t^*]} = k, \boldsymbol{m}_i, \{\boldsymbol{x}_i^{[t]}\}_{t=1}^{t^*}) \\ =&\sigma\left((\boldsymbol{z}_i^{[t^*]})^T \boldsymbol{\beta}_{k,l^k,r}\right), \end{aligned} \tag{8}$$

with $r = 1$ in all cases, because a Bernoulli variable has an one-dimensional outcome.

The product of Eq. (6) and (7) as well as that of Eq. (6) and (8) yields the joint probability of both therapy feature and category as

$$\mathbb{P}(F_{i,k,l^k}^{[t^*]} = r \wedge C_i^{[t^*]} = k \mid \boldsymbol{m}_i, \{\boldsymbol{x}_i^{[t]}\}_{t=1}^{t^*}). \tag{9}$$

But due to the fact that

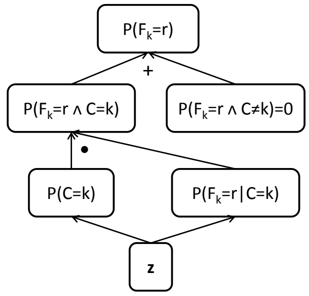
$$\mathbb{P}(F_{i,k,l^k}^{[t^*]} = r \wedge C_i^{[t^*]} \neq k \mid \boldsymbol{m}_i, \{\boldsymbol{x}_i^{[t]}\}_{t=1}^{t^*}) = 0, \tag{10}$$

in all cases, this joint probability of Eq. (9) is equal to

$$\mathbb{P}(F_{i,k,l^k}^{[t^*]} = r | \boldsymbol{m}_i, \{\boldsymbol{x}_i^{[t]}\}_{t=1}^{t^*}), \tag{11}$$

applying the law of total probability, yielding the marginal prediction and allowing us to perform the optimization against the target vector. The calculation with these probabilities is illustrated in Fig. 4. A same design can also be found in [14], where they factorize a large softmax layer into such a tree-like hierarchy.

Figure 4. A simplified illustration of deriving the marginal probability of the therapy feature. From the vector $\boldsymbol{z}$ representing a patient one calculates the category probability and the feature probability conditioned on the category. Then product of the two yields the joint probability of feature and category. Combined with the joint probability of feature and non-category, which is always 0, one would get the marginal probability of the feature.



In [13] a very similar approach is referred to as the Multinomial Model with Hierarchically Structured Response. The major difference lies in the fact that in [13] only one multinomial response on the second level is linked with each category on the first level. This is apparently not sufficient for our data situation where multiple multinomial therapy features fall into each therapy category. Therefore we extend this model and allow for multiple of such links.

Finally we illustrate the complete model architecture in Fig. 5. There the RNN encoder outputs its last hidden state that represents the whole sequence and is concatenated with the latent representation mapped from the static patient information. This concatenated vector forms the input to the hierarchical model, which in the first step calculates the therapy category probabilities and in the second step the therapy feature probabilities conditioned on corresponding category. These two levels of probabilities are multiplied, giving the joint probabilities of category and feature, which are equivalent to marginal feature probabilities as proven in Eq. (10).

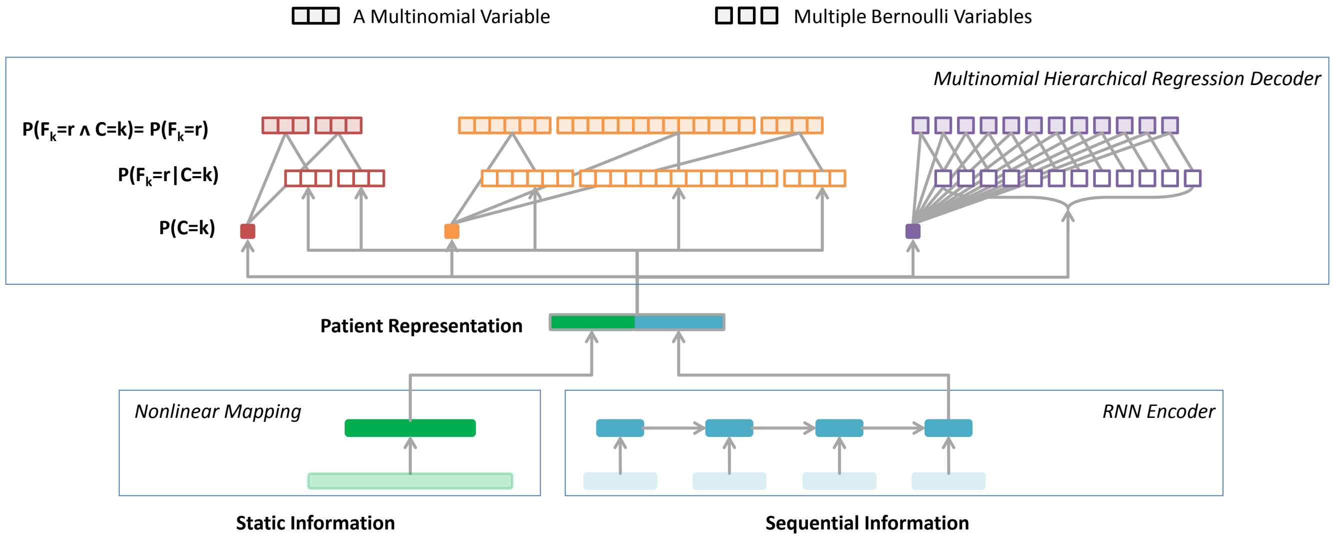
5. Experiment

We evaluate our encoder-decoder model from two aspects. First we assess the prediction quality and then demonstrate that our model can be exploited to identify similar historical patient cases in a very efficient way.

5.1. Modeling of Therapy Decisions

In order to take into account the prediction stability, we conduct cross-validation by splitting the 2,869 patients into 5 disjoint sets, and then query their corresponding sequences to form the training and test sets. In contrast to performing the splitting on the level of sequences, this approach guarantees

Figure 5. Our proposed model architecture. The radiotherapy features consist of two 3D multinomial variables (red-colored). The systemic therapies consist of on 4D, one 16D and one 3D multinomial variables (orange-colored). The surgery feature consists of 10 Bernoulli variables (purple-colored).



that the model only predicts for completely new patients whose information was never –not even partially– present during training, making the experiments more challenging and realistic. For the rest of this section we report the average performances of cross-validations, for all experimental settings including baseline models.

With respect to the sizes of $\boldsymbol{q}_i$ that represents the static information and $\boldsymbol{h}_i^{[t^*]}$ that represents the medical history, we conduct experiments with two settings. In a smaller setting we define $\boldsymbol{q}_i \in \mathbb{R}^{64}$ and $\boldsymbol{h}_i^{[t^*]} \in \mathbb{R}^{128}$ and present the results in Tab. 5, while Tab. 6 provides experimental results with a larger setting of $\boldsymbol{q}_i \in \mathbb{R}^{128}$ and $\boldsymbol{h}_i^{[t^*]} \in \mathbb{R}^{256}$.

Further hyper-parameters are set as follows: the output of $\boldsymbol{h}_i^{[t^*]}$ in RNN is activated with $\tanh$. We apply 0.25 Dropout [23] for weights in RNNs and 0.001 Ridge penalization for the MHR and logistic decoders. Each model instance is trained with Adam [24] step rule for a maximum of 1000 iterations with early stopping mechanism.

We present two classes of evaluation metrics. First, column-wise average Area Under ROC (AUROC) and Area Under Precision-Recall-Curve (AUPRC), which are well-known metrics applied to measure the classification quality, should indicate the models' capability to assign patients to the correct therapy features. Secondly, we report multi-label ranking-based metrics of Coverage Error (CE) [25] and Label Ranking Average Precision (LRAP) [26] in the scikit-learn library [27]. In contrast to precision and recall based metrics, they are calculated row-wise and thus evaluate for each patient how many recommended therapies were actually prescribed. LRAP ranges between 0 and 1 just as AUROC and AUPRC. CE describes how many steps one has to go in a ranked list of recommendations till one covers all ground truth labels. In our case, the average number of labels in each patient case is 4.4 and the total number of possible labels is 39. The CE shall therefore be ideally 4.4, suggesting a perfect prediction, and be 39 in worst case scenario (Tab. 4).

Table 4. Results of experiments with two weak baselines: Random Prediction and Constant Most Popular Prediction.

Weak Baselines	AUROC	AUPRC	CE	LRAP
Random	49.7%	9.4%	38.2	11.2%
Most Popular	50.0%	21.3%	13.9	38.6%

We experiment with three encoders and two decoders.

The baseline encoder is a simple Feed-Forward Layer (FFL) consuming the raw sequential information that is aggregated with respect to time. Then the aggregated feature vector is concatenated with the static feature vector for each patient case. Such aggregation can be interpreted as a hand-engineered feature processing, where each feature represents the total number of observed feature values. It also corresponds to the bag-of-words approach [28] in Natural Language modeling, this approach completely neglects the *order* in which the feature values are observed. As a more advanced solution we apply GRU and LSTM as RNN encoders as introduced in Section 4.1, which are expected to capture the information regarding the events order as well.

The baseline decoder is a single-layered logistic regression, which is a popular choice in multi-class multi-label classification tasks in machine learning. Please note that this approach does not fully satisfy the distribution assumption of the target. For instance, a therapy *feature* variable is multinomially distributed, implying the mutual exclusiveness of the probable outcomes of the feature values and this aspect cannot be taken into account with a flat logistic regression. Such mutual exclusiveness has to be taken into account especially in clinical data. For instance a physician is only supposed to prescribe one medication from a class of related medications. Since our proposed MHR model, presented in

Sec. 4.2, is mathematically solid from this perspective, it is interesting to see it actually outperforms a more popular but less accurate alternative.

We conduct experiments applying all possible combinations of encoders and decoders, to identify i) which combination yields the best prediction performance and ii) which encoder contributes the most to the model performances given the same decoder, and vice versa.

Table 5. Average Results of Experiments with Different Encoders and Decoders, with $\boldsymbol{q}_i \in \mathbb{R}^{128}$ and $\boldsymbol{h}_i^{[t^*]} \in \mathbb{R}^{256}$

Encoder	Decoder	AUROC	AUPRC	CE	LRAP
FFL	Logistic	69.4%	13.4%	12.61	48.6%
	MHR	70.3%	13.9%	11.79	49.3%
GRU	Logistic	81.8%	28.8%	8.57	61.3%
	MHR	**82.1%**	**31.2%**	**8.26**	**62.3%**
LSTM	Logistic	79.6%	24.7%	9.47	57.9%
	MHR	81.9%	30.2%	8.53	61.4%

Table 6. Average Results of Experiments with Different Encoders and Decoders, with $\boldsymbol{q}_i \in \mathbb{R}^{64}$ and $\boldsymbol{h}_i^{[t^*]} \in \mathbb{R}^{128}$

Encoder	Decoder	AUROC	AUPRC	CE	LRAP
FFL	Logistic	69.8%	13.4%	12.83	48.3%
	MHR	70.2%	13.9%	11.83	49.2%
GRU	Logistic	80.0%	26.2%	9.28	59.0%
	MHR	**81.3%**	**28.2%**	**8.71**	**61.3%**
LSTM	Logistic	78.7%	23.0%	9.93	56.4%
	MHR	80.6%	26.7%	9.12	59.5%

Comparing Tab. 5 with 6 one could observe that, with a larger size for the representation vector, the prediction quality can be improved in almost all cases. With both parameter settings the combination of GRU encoder and the hierarchical decoder yields the best quality scores.

It is to note that both decoders on top of the baseline FFL encoder show suboptimal results compared with those on top of RNN encoders, i.e., GRU and LSTM encoders significantly boost the prediction quality even with a mere logistic regression as decoder. On the other hand, the MHR model further improves the prediction quality in comparison with a flat logistic regression. This suggests that the RNN encoders contribute a larger proportion to the prediction quality, while the multinomial hierarchical decoder alone does not improve the model to a significant extent without a decent encoder model. One could draw the conclusion that the encoded representation of a patient case plays a central role in this model. In total, the prediction best quality is provided by GRU encoder and MHR decoder.

5.2. Identification of Similar Patient Cases

In a realistic application scenario in a clinic, it is as important to provide physicians with recommended therapies as to provide a list of similar patient cases. If the set of similar patient cases have received therapies similar to the recommended ones, it will support these recommendations and encourage the physician to interpret the recommendations with more confidence. But due to the fact that patients have medical histories of variable lengths, it is nontrivial to apply common distance metrics directly on the patient features to quantify the similarity. For instance, it is impossible to mathematically directly calculate the distance between a patient having undergone a breast conservation surgery and another patient with one mastectomy followed by three successive radiotherapies, although it might be obvious for a physician to tell the difference/similarity.

To this end, we propose that the derived latent vector of $\boldsymbol{z}_i^{[t^*]}$, representing the patient i's profile up to time t^*, can be exploited to identify similar patient cases, since all such vectors have the same dimension.

Using a trained encoder network, we map all training patient cases and the a test case into the latent feature space and define there a k-NN model. The k training cases neighboring the test case can therefore provide a prediction for the test case. This approach reusing trained representation is closely related to the so-called transfer learning [29], where one exploits the latent representations learned for one task for new tasks. In our case, however, we have the same task solved by a new predictive model that consumes the learned representations.

If the k-NN model is able to identify patient cases having received therapies that agree with the recommended ones, then the latent vectors correctly represent such similarity.

We report the results of such k-NN models that is applied on latent representations originally learned with GRU and LSTM encoder combined with logistic and hierarchical decoder in Tab. 7 and 8 for the two parameter settings of the latent vector sizes, respectively. Please note that an RNN encoder converges to different weight parameters combined with different decoders. Therefore, although we only apply the encoder network, it is necessary to differentiate between the two decoder cases. The k is here set to be 30. We realize that a smaller k would hurt the prediction quality and a larger k does not further improve the model.

Table 7. Results of experiments with k-NN on top of the latent representations derived by different model architecture settings, with $\boldsymbol{q}_i \in \mathbb{R}^{128}$ and $\boldsymbol{h}_i^{[t^*]} \in \mathbb{R}^{256}$.

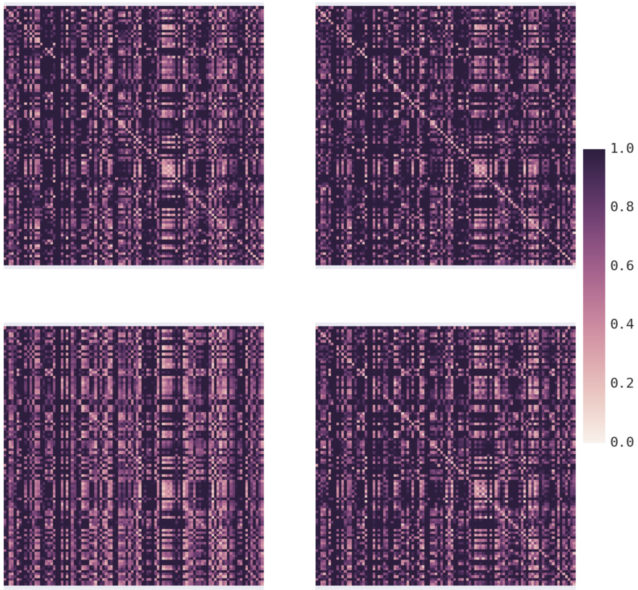
orig. Encoder	orig.Decoder	AUROC	AUPRC	CE	LRAP
GRU	Logistic	78.7%	30.0%	**9.69**	63.0%
	MHR	**79.3%**	**32.2%**	9.82	63.2%
LSTM	Logistic	78.6%	28.7%	9.80	62.7%
	MHR	79.0%	**32.2%**	9.83	**63.3** %

Table 8. Results of experiments with kNN on top of the latent representations derived by different model architecture settings, with $\boldsymbol{q}_i \in \mathbb{R}^{64}$ and $\boldsymbol{h}_i^{[t^*]} \in \mathbb{R}^{128}$

orig. Encoder	orig.Decoder	AUROC	AUPRC	CE	LRAP
GRU	Logistic	78.1%	28.6%	9.75	62.9%
	MHR	**79.1%**	29.8%	9.71	**63.3%**
LSTM	Logistic	78.3%	28.5%	9.78	62.5%
	MHR	**79.1%**	**30.2%**	**9.69**	63.2%

One could observe that the k-NN performances are in total quite close to those reported in Tab. 5 and 6.

Figure 6. The distance matrix between a sample of 100 decisions predicted by i) the encoder-decoder model and ii) the k-NN model based on the latent representations with four different encoder-decoder settings. Top-left: GRU+sigmoid; top-right: GRU+hierarchical; bottom-left: LSTM+sigmoid; bottom-right: LSTM+hierarchical.



However, if we build the same k-NN classifier on top of the raw features aggregated in time and concatenated with the static features, the prediction quality is observed to be much worse: The AUROC and AUPRC decrease to 75.2% and 23.3%, respectively, while the CE and LRAP to 11.61 and 56.8%, respectively. This suggests that the RNN encoder is capable of generating more dense and informative latent features for each patient case.

In order to compare the concrete decision made by the original encoder-decoder model and the k-NN model on each specific patient case, we calculate the Euclidean distances between the predictions made by i) the encoder-decoder model and ii) the k-NN model based on the same encoder and decoder setting for each patient case. We visualize in Fig. 6 the distance matrices as heat map. One could observe that the diagonal entries, which represent prediction distances between the complete model and the k-NN model for the same patient cases, are systematically lower in value.

To this end, we argue that the latent vectors can represent a patient case with medical history of variable length in a unified vector space, where the topological characteristics of patients are well preserved. This additional feature of our encoder-decoder model enables the identification of similar patient cases, which can support and supplement the predictive recommendations.

6. Conclusion

We have proposed an Encoder-Decoder network that predicts physicians' therapy decisions as well as provides a list of similar patient cases. The model consists of an RNN encoder that learns an abstract representation of the patient profile, and a hierarchical regression that decodes the latent representation into therapy predictions. Such a predictive model can serve to support clinical decisions, to detect anomalous prescriptions and to support physician by searching for similar historical cases.

We have conducted experiments on a large real-world dataset collected from almost three thousands of metastatic breast cancer patients. The experimental results demonstrate that the RNN encoder greatly improves the modeling quality compared with plain feed-forward models that consume aggregated sequential features. The hierarchical regression model also outperforms a flat logistic regression as a decoder. We have also shown that our model is capable of providing lists of similar patient cases, although it is nontrivial to measure distance among patients, when they all have medical histories of variable lengths.

The generic contribution of this work consists of following aspects:

- We transfer the popular Encoder-Decoder architecture from NLP to the clinical domain;
- We propose a hierarchical classifier that mimics the actual multi-step decision procedure;
- We empirically prove that the latent vector representing each patient case produced by RNN encoders in general facilitates the prediction with k-NN, logistic regression and MHR;
- We showed that such latent representations can be exploited to identify similar patients with higher quality than with aggregated sequential features.

Encouraged by the success of the RNN models in handling sequential data, one interesting and realistic improvement of the model would be to integrate attention mechanisms [30, 31] into the RNN encoder. The model would, for instance, be able to identify which historical event has contributed most to the decision, which could further improve the model's interpretability and encourage its application in clinics.

References

[1] V. Tresp, M. Overhage, M. Bundschus, S. Rabizadeh, P. Fasching, and S. Yu, "Going digital: A survey on digitalization and large scale data analytics in healthcare," *arXiv preprint arXiv:1606.08075*, 2016.

[2] R. Rahman and C. K. Reddy, "Electronic health records: a survey," *Healthcare Data Analytics*, vol. 36, p. 21, 2015.

[3] E. Choi, M. T. Bahadori, and J. Sun, "Doctor ai: Predicting clinical events via recurrent neural networks," *arXiv preprint arXiv:1511.05942*, 2015.

[4] C. Esteban, D. Schmidt, D. Krompaß, and V. Tresp, "Predicting sequences of clinical events by using a personalized temporal latent embedding model," in *Healthcare Informatics (ICHI), 2015 International Conference on*. IEEE, 2015, pp. 130–139.

[5] C. Esteban, O. Staeck, Y. Yang, and V. Tresp, "Predicting clinical events by combining static and dynamic information using recurrent neural networks," *arXiv preprint arXiv:1602.02685*, 2016.

[6] M. A. Brookhart, T. Stürmer, R. J. Glynn, J. Rassen, and S. Schneeweiss, "Confounding control in healthcare database research: challenges and potential approaches," *Medical care*, vol. 48, no. 6 0, p. S114, 2010.

[7] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[8] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A neural probabilistic language model," *journal of machine learning research*, vol. 3, no. Feb, pp. 1137–1155, 2003.

[9] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzell, "Learning to diagnose with lstm recurrent neural networks," *arXiv preprint arXiv:1511.03677*, 2015.

[10] E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, and W. Stewart, "Retain: An interpretable predictive model for healthcare using reverse time attention mechanism," in *Advances in Neural Information Processing Systems*, 2016, pp. 3504–3512.

[11] J. Koutnik, K. Greff, F. Gomez, and J. Schmidhuber, "A clockwork rnn," *arXiv preprint arXiv:1402.3511*, 2014.

[12] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 2011, pp. 142–150.

[13] G. Tutz, *Regression for Categorical Data:*, ser. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2011.

[14] F. Morin and Y. Bengio, "Hierarchical probabilistic neural network language model." in *Aistats*, vol. 5. Citeseer, 2005, pp. 246–252.

[15] P. Kaatsch, C. Spix, S. Hentschel, A. Katalinic, S. Luttmann, C. Stegmaier, S. Caspritz, J. Cernaj, A. Ernst, J. Folkerts *et al.*, "Krebs in deutschland 2009/2010," 2013.

[16] C. Rauh and W. Matthias, "Interdisziplinäre s3-leitlinie für die diagnostik, therapie und nachsorge des mammakarzinoms," 2008.

[17] P. Fasching, S. Brucker, T. Fehm, F. Overkamp, W. Janni, M. Wallwiener, P. Hadji, E. Belleville, L. Häberle, F. Taran, D. Luftner, M. Lux, J. Ettl, V. Muller, H. Tesch, D. Wallwiener, and A. Schneeweiss, "Biomarkers in patients with metastatic breast cancer and the praegnant study network," *Geburtshilfe Frauenheilkunde*, vol. 75, no. 01, pp. 41–50, 2015. [Online]. Available: http://www.praegnant.org/

[18] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[19] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[20] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with lstm," *Neural computation*, vol. 12, no. 10, pp. 2451–2471, 2000.

[21] Y. Kim, Y. Jernite, D. Sontag, and A. M. Rush, "Character-aware neural language models," *arXiv preprint arXiv:1508.06615*, 2015.

[22] T. Mikolov, M. Karafiát, L. Burget, J. Cernockỳ, and S. Khudanpur, "Recurrent neural network based language model." in *Interspeech*, vol. 2, 2010, p. 3.

[23] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting." *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[24] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[25] G. Tsoumakas, I. Katakis, and I. Vlahavas, "Mining multi-label data," in *Data mining and knowledge discovery handbook*. Springer, 2009, pp. 667–685.

[26] G. Madjarov, D. Kocev, D. Gjorgjevikj, and S. Džeroski, "An extensive experimental comparison of methods for multi-label learning," *Pattern Recognition*, vol. 45, no. 9, pp. 3084–3104, 2012.

[27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[28] Z. S. Harris, "Distributional structure," *Word*, vol. 10, no. 2-3, pp. 146–162, 1954.

[29] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[30] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in Neural Information Processing Systems*, 2015, pp. 577–585.

[31] V. Mnih, N. Heess, A. Graves *et al.*, "Recurrent models of visual attention," in *Advances in neural information processing systems*, 2014, pp. 2204–2212.