Learning Individualized Treatment Rules With Estimated Translated Inverse Propensity Score

Zhiliang Wu[1,2], Yinchong Yang[1], Yunpu Ma[1,2], Yushan Liu[1,2], Rui Zhao[1,2], Michael Moor[3], Volker Tresp[1,2]
[1]Siemens AG, [2]LMU Munich, [3]ETH Zurich

IEEE International Conference on Healthcare Informatics, 2020, Best Paper Award

IEEE ICHI 2020

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

SIEMENS

Agenda

- Motivation
- Methods
- Experiments
- Summary and Outlook

Motivation
Predictive Modelling vs. Individualized Treatment Rules

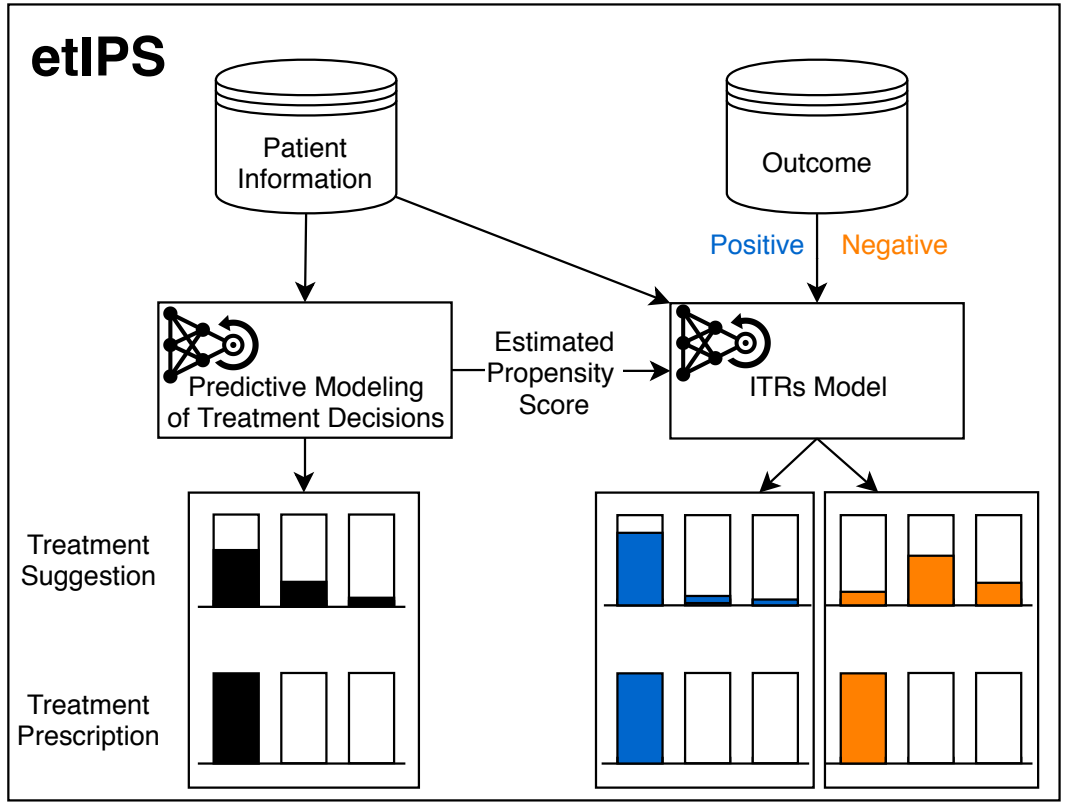
Predictive modelling of treatment decisions

[Esteban 2016, Yang 2017]

- Observed treatment decisions as targets
- Propensity score estimation
- No outcome information involved

Learning of the Individualized Treatment Rules (ITRs)

- A treatment policy that is expected to generate a better outcome for an individual patient
- Proposed framework: etIPS
- Integration of the observed outcome information with predictive modelling



Esteban, C., Staeck, O., Baier, S., Yang, Y., & Tresp, V. (2016, October). Predicting clinical events by combining static and dynamic information using recurrent neural networks. In 2016 IEEE International Conference on Healthcare Informatics (ICHI) (pp. 93-101). IEEE.

Yang, Y., Fasching, P. A., & Tresp, V. (2017, August). Predictive modeling of therapy decisions in metastatic breast cancer with recurrent neural network encoder and multinomial hierarchical regression decoder. In 2017 IEEE International Conference on Healthcare Informatics (ICHI) (pp. 46-55). IEEE.

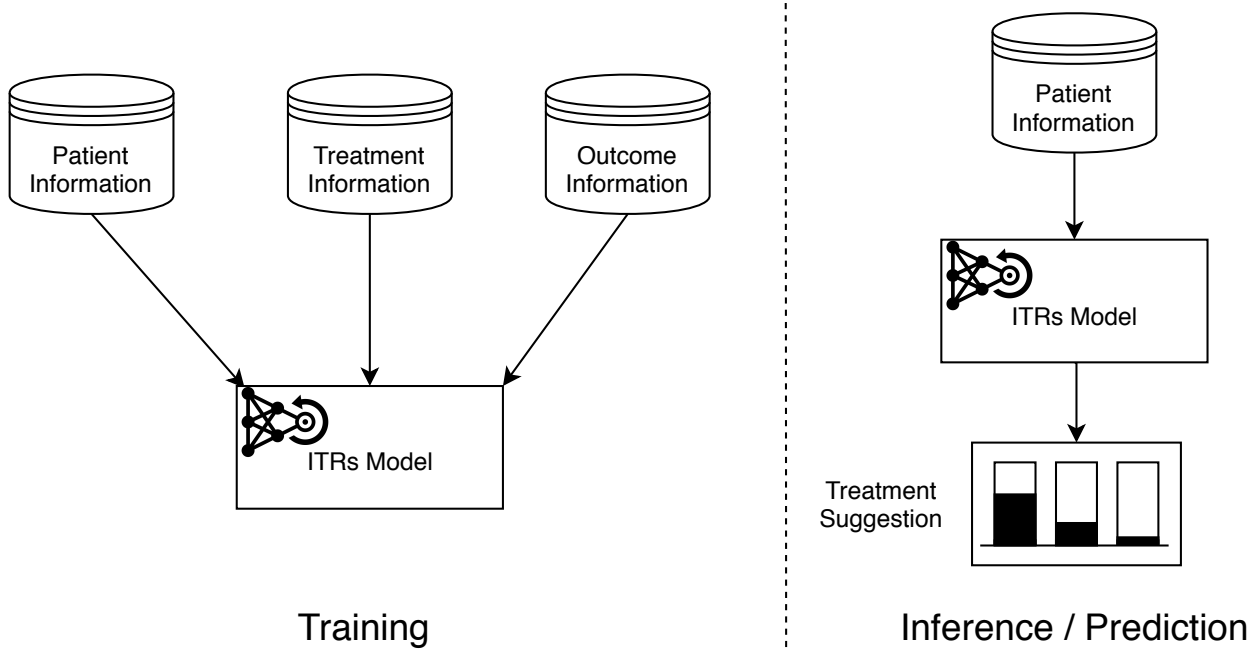
Motivation
Problem Settings of ITRs

Training

- Treatment policy optimization
- Batch learning setup: No new samples collected
- Contextual bandits: Batch Learning from Bandit Feedback (BLBF)

Inference / Prediction

- The same as predictive modelling of treatment decisions



Methods
Propensity Score-Based Training Objective

X_i: covariate of a patient
a_i: treatment
δ_i: loss

- Find a new policy $\pi_{\boldsymbol{w}}$ that minimizes the risk

$$r\left(\pi_{\boldsymbol{w}}\right) = \mathbb{E}_{X\sim\mathbb{P}(X)}\mathbb{E}_{a\sim\pi_w(a|X)}[\delta(X,a)]$$
$$= \mathbb{E}_{X\sim\mathbb{P}(X)}\mathbb{E}_{a\sim\mathbb{P}(a|X)}\left[\delta(X,a)\cdot\frac{\pi_{\boldsymbol{w}}(a|X)}{\mathbb{P}(a|X)}\right]$$

	Propensity score overfitting [Swaminathan2015]	mini-batch SGD friendly
Inverse Propensity Score Estimator (IPS) $\hat{r}_{\mathrm{IPS}}\left(\pi_w\right) = \frac{1}{n}\sum_{i=1}^{n}\delta_i\frac{\pi_w\left(a_i \mid X_i\right)}{\mathbb{P}\left(a_i \mid X_i\right)}$	✗	✓
Self-Normalized IPS Estimator (SNIPS) [Swaminathan 2015] $\hat{r}_{\mathrm{SNIPS}}\left(\pi_{\boldsymbol{w}}\right) = \frac{\frac{1}{n}\sum_{i=1}^{n}\delta_i\frac{\pi_w(a_i\|X_i)}{\mathbb{P}(a_i\|X_i)}}{\frac{1}{n}\sum_{j=1}^{n}\frac{\pi_w(a_j\|X_j)}{\mathbb{P}(a_j\|X_j)}}$	✓	✗
λ-translated IPS estimator [Joachims2018] $\hat{r}^{\lambda}_{\mathrm{IPS}}\left(\pi_w\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\delta_i-\lambda\right)\frac{\pi_w\left(a_i \mid X_i\right)}{\mathbb{P}\left(a_i \mid X_i\right)}$	✓	✓

Swaminathan, Adith, and Thorsten Joachims. "The self-normalized estimator for counterfactual learning." *advances in neural information processing systems*. 2015.
Joachims, Thorsten, Adith Swaminathan, and Maarten de Rijke. "Deep learning with logged bandit feedback." *International Conference on Learning Representations*. 2018.

Methods
Intuition of λ-translated IPS Estimator

The gradient of trainable parameters $\boldsymbol{w}$ in λ-translated IPS estimator

$$\hat{r}^{\lambda}_{\mathrm{IPS}}(\pi_{\boldsymbol{w}}) = \frac{1}{m}\sum_{i=1}^{m}(\delta_i - \lambda)\frac{\pi_{\boldsymbol{w}}(a_i|X_i)}{\mathbb{P}(a_i|X_i)} \Longrightarrow \nabla\hat{r}^{\lambda}_{\mathrm{IPS}} = \frac{1}{m}\sum_{i=1}^{m}\boxed{\delta_i - \lambda}{\mathbb{P}(a_i|X_i)}\nabla\pi_{\boldsymbol{w}}(a_i|X_i)$$

For loss δ_i

- 0 for treatments with positive outcome
- 1 for treatments with negative outcome

Two extremes of λ-translated IPS estimator

- $\lambda = 0$: only the gradient of negative outcome is received by $\boldsymbol{w}$. → a completely different policy from physicians
- $\lambda = 1$: only the gradient of positive outcome is received by $\boldsymbol{w}$. → a very similar policy to the physicians
- $\lambda \in (0, 1)$: gradient of both negative and positive outcome can be received by $\boldsymbol{w}$.
- The optimal λ^*: found through grid search.

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine learning, 8(3), 229-256.

Methods
Estimated Translated Inverse Propensity Score

Algorithm 1: etIPS

Input: A dataset of the form $\{X_i, a_i, \delta_i\}_{i=1}^m$.

Output: The policy of the optimal ITRs $\pi_{\boldsymbol{w}^*}(a|X)$.

Learn the physicians' policy $\hat{\mathbb{P}}(a|X)$ with $\{X_i, a_i\}_{i=1}^m$ using the network structure in Fig. 3.

Compute the estimated propensity score $\hat{p}_i := \hat{\mathbb{P}}(a = a_i|X_i)$ for all i.

for $\lambda_j \in (0, 1)$ **do**

$$\boldsymbol{w}_j^* \leftarrow \arg\min_{\boldsymbol{w}_j} \left\{ \frac{1}{m} \sum_{i=1}^m (\delta_i - \lambda_j) \frac{\pi_{\boldsymbol{w}_j}(a_i|X_i)}{\hat{p}_i} \right\}$$

$$s_j \leftarrow \frac{1}{m} \sum_{i=1}^m \frac{\pi_{\boldsymbol{w}_j^*}(a_i|X_i)}{\hat{p}_i}$$

end

$$s^*, \boldsymbol{w}^* \leftarrow \arg\min_{s_j, \boldsymbol{w}_j^*} \left\{ \frac{1}{s_j} \frac{1}{m} \sum_{i=1}^m \delta_i \frac{\pi_{\boldsymbol{w}_j^*}(a_i|X_i)}{\hat{p}_i} \right\}$$

return $\pi_{\boldsymbol{w}^*}(a|X)$

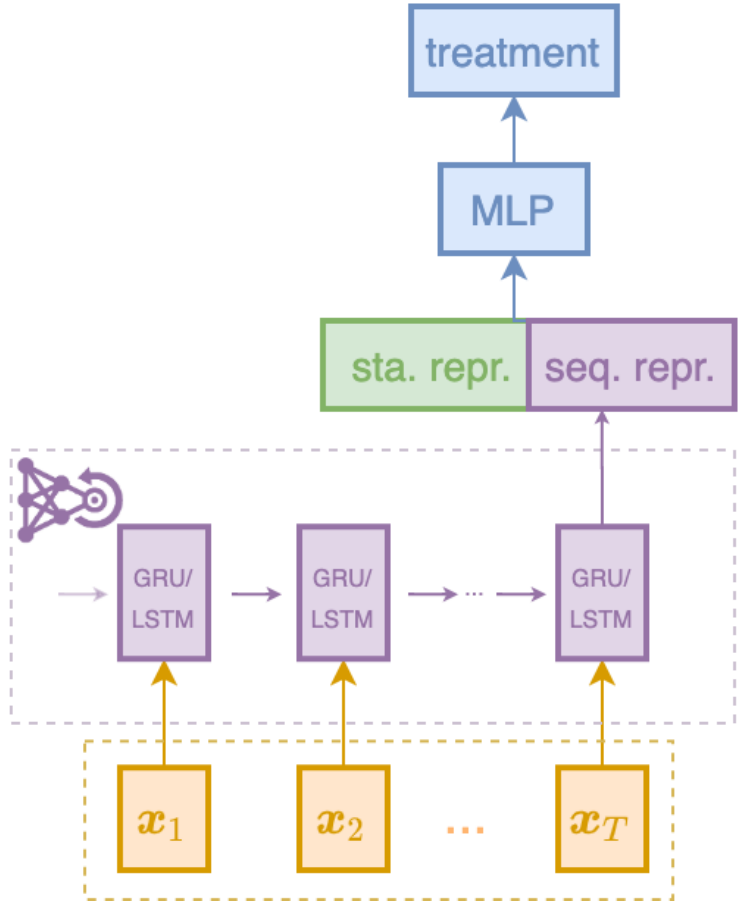
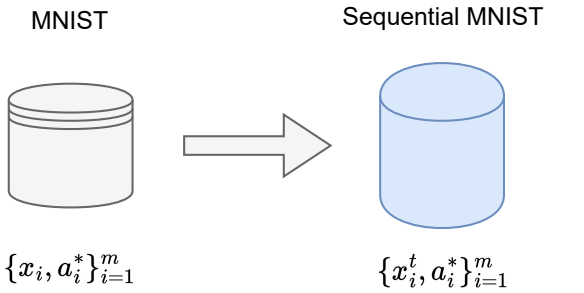


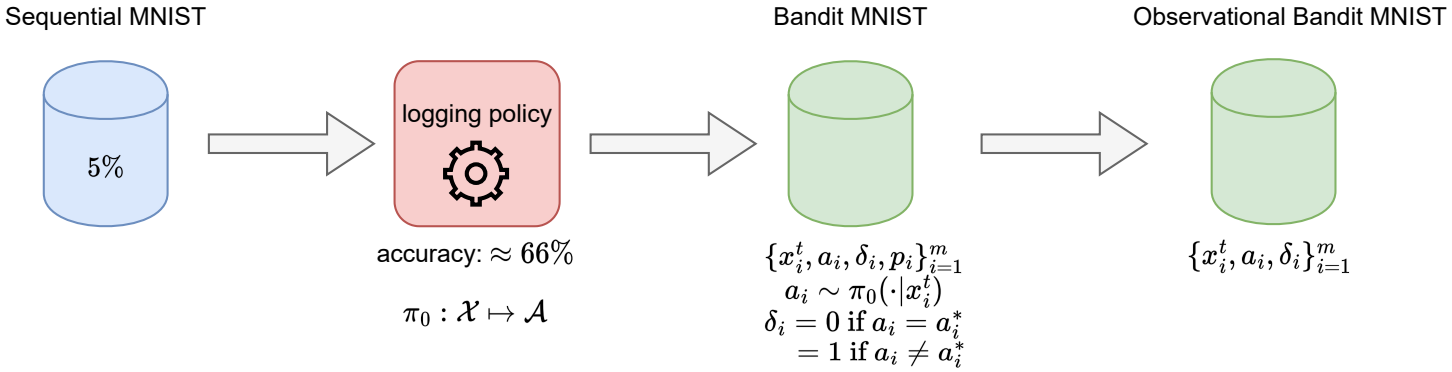
Fig. 3 Network architecture, first proposed in [Yang 2017]

Yang, Y., Fasching, P. A., & Tresp, V. (2017, August). Predictive modeling of therapy decisions in metastatic breast cancer with recurrent neural network encoder and multinomial hierarchical regression decoder. In 2017 IEEE International Conference on Healthcare Informatics (ICHI) (pp. 46-55). IEEE.

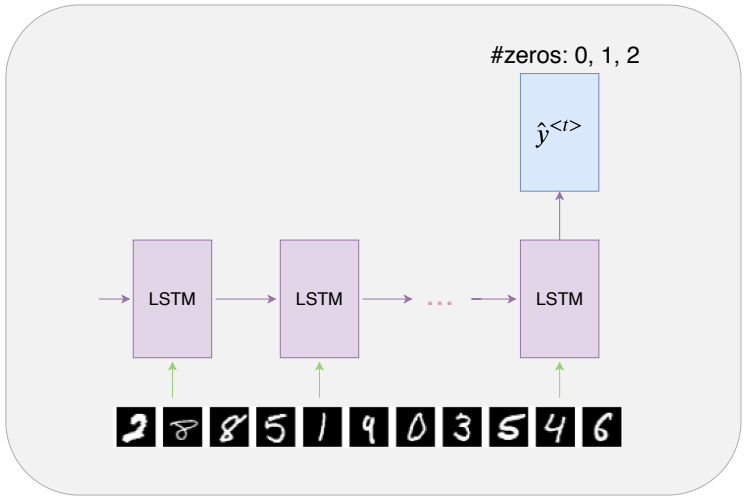
Experiments
Simulation Study



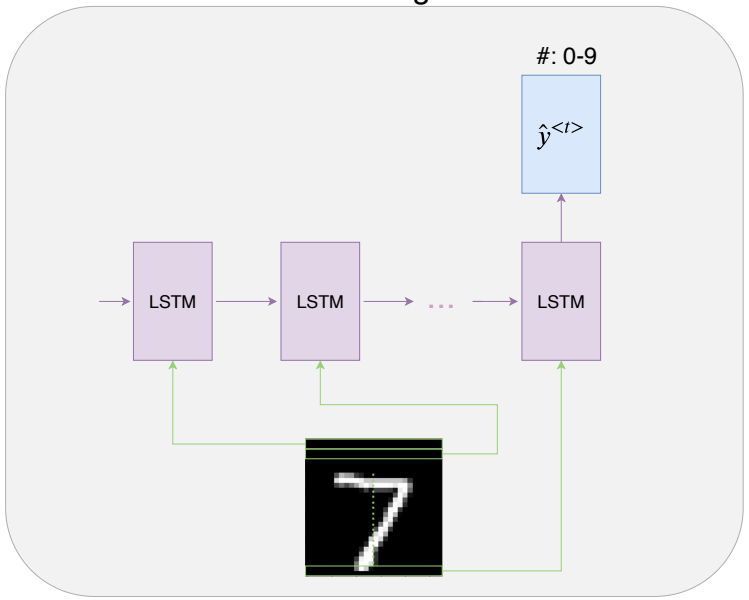
- Context: A sequence of images
- Action: Label prediction of the given sequence
- Loss: Correctness of the prediction
- Propensity: Probability of the respective label prediction



Supervised to Bandit Conversion Method [Agarwal 2014]



zero counting MNIST



row-by-row MNIST

Agarwal, Alekh, et al. "Taming the monster: A fast and simple algorithm for contextual bandits." *International Conference on Machine Learning*. PMLR, 2014.

Experiments
Simulation Study

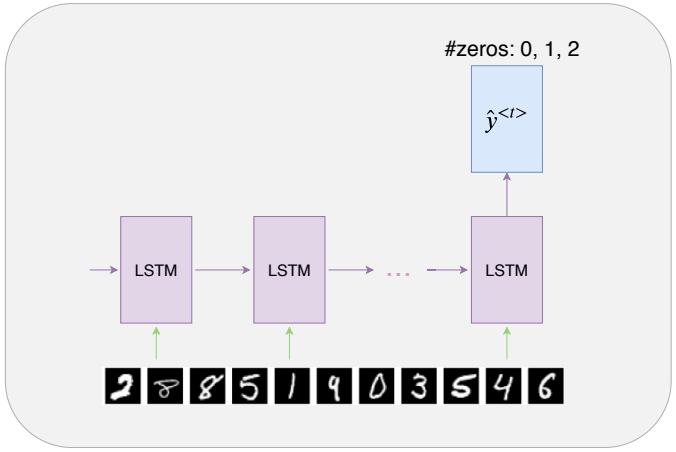
Hypothesis validation

- Without ground truth labels, the propensity score-based objective function is applicable to classification tasks with sequential data.
- The estimated propensity score could be used to replace the true propensity score in the objective function. *(cf. MLIPS in [Xie 2018])*

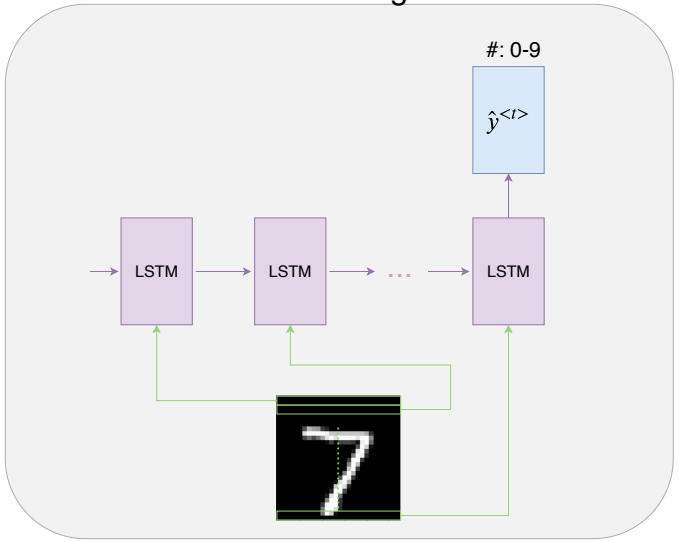
TABLE I
ACCURACY OF DIFFERENT APPROACHES ON SEQUENTIAL CLASSIFICATION TASKS

	propensity score	zeros counting MNIST	row-by-row MNIST
DM	-*	0.343 ± 0.0001	0.098 ± 0.0022
RP	-*	0.363 ± 0.0001	0.103 ± 0.0001
IPS	true	0.301 ± 0.012	0.020 ± 0.0061
tIPS	true	0.899 ± 0.0229	0.931 ± 0.0852
eIPS	estimated	0.319 ± 0.0075	0.016 ± 0.0098
etIPS	estimated	$\mathbf{0.923 \pm 0.0122}$	$\mathbf{0.953 \pm 0.0390}$

*The propensity score is not involved in the algorithm.



zero counting MNIST



row-by-row MNIST

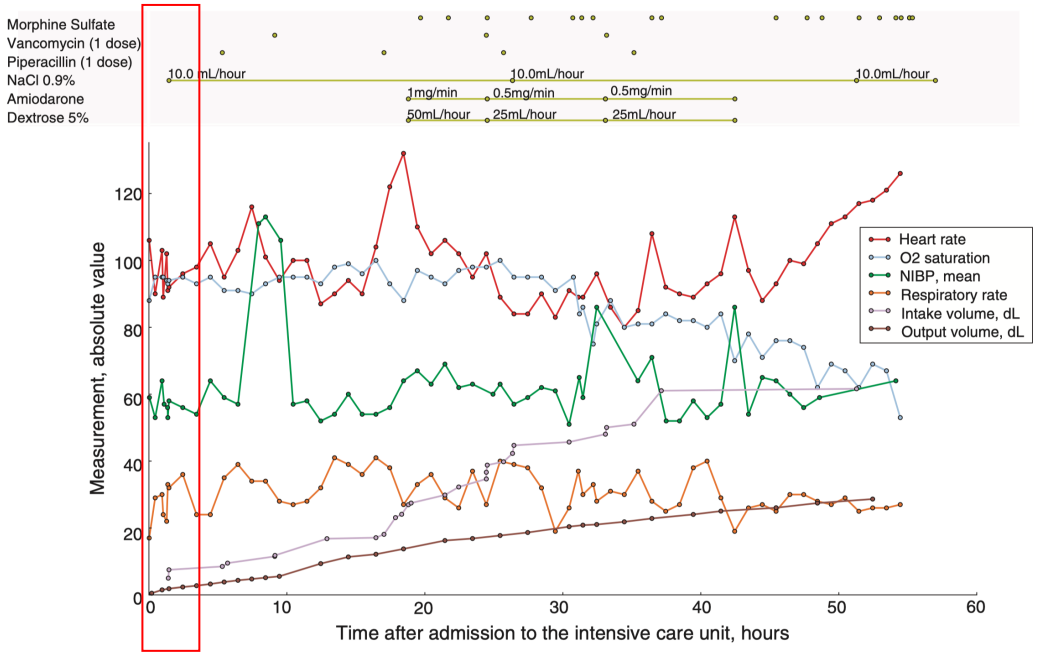
Xie, Y., Liu, B., Liu, Q., Wang, Z., Zhou, Y., & Peng, J. (2018, September). Off-Policy Evaluation and Learning from Logged Bandit Feedback: Error Reduction via Surrogate Policy. In International Conference on Learning Representations.

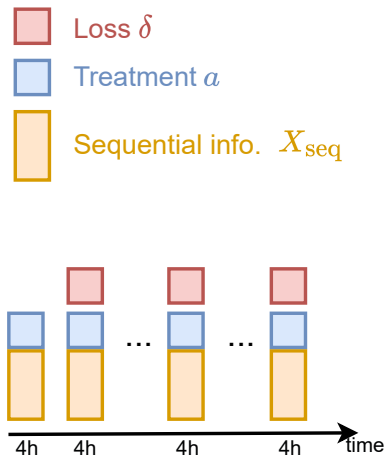
Experiments
MIMIC-III

Medical Information Mart for Intensive Care database (MIMIC-III)

- A freely accessible database
- 53,423 Intensive Care Unit (ICU) admissions

Use case from [Komorowski 2018]





Dosage actions

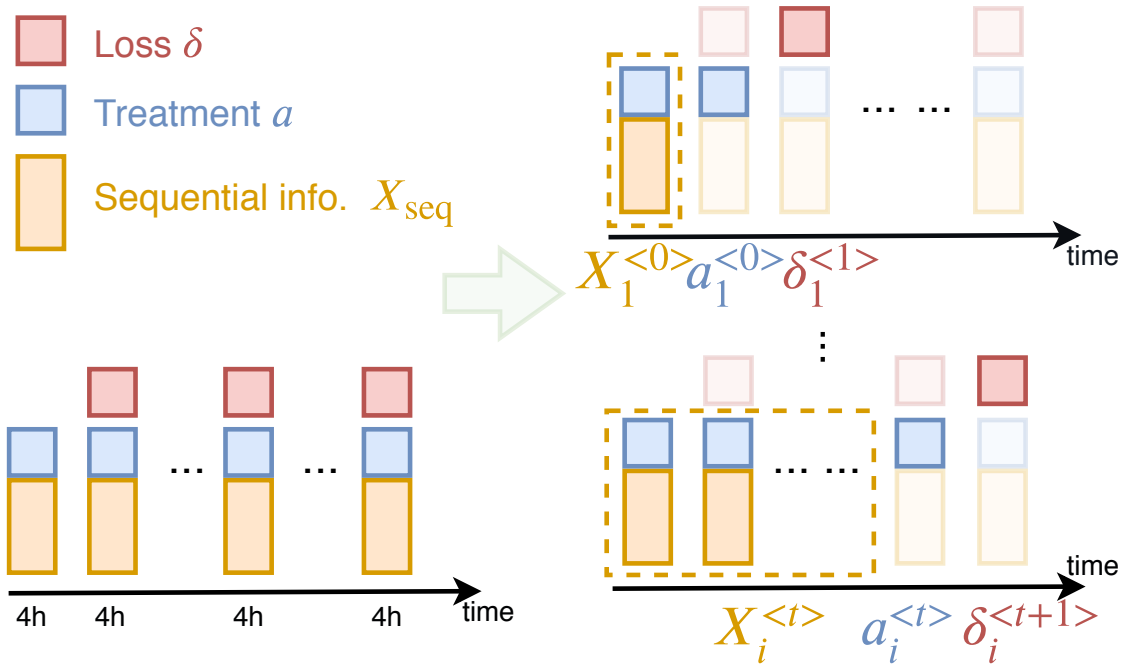
Dose of i.v. fluid \ Dose of vasopressor	1	2	3	4	5
1	1	2	3	4	5
2	6	7	8	9	10
3	11	12	13	14	15
4	16	17	18	19	20
5	21	22	23	24	25

Komorowski, M., Celi, L. A., Badawi, O., Gordon, A. C., & Faisal, A. A. (2018). The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. Nature medicine, 24(11), 1716-1720.

Experiments
MIMIC-III

Use case from [Komorowski 2018]

- A cohort of patients fulfilling the Sepsis-3 criteria [1]
- 20,944 admissions
- 224,333 samples



Komorowski, M., Celi, L. A., Badawi, O., Gordon, A. C., & Faisal, A. A. (2018). The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. Nature medicine, 24(11), 1716-1720.
[1] https://github.com/matthieukomorowski/AI_Clinician

Experiments
MIMIC-III

Different evaluation metrics to indicate the expected risk (lower is better)

- Average Treatment Effects under the new policy (ATENP) [Zhou 2017, Komorowski 2018]
- Doubly Robust Estimator (DR) [Dudik 2011]

TABLE II
EVALUATION WITH DIFFERENT RISK ESTIMATORS

	ATENP	DR
Predictive Modeling	-0.019 ± 0.0021	0.523 ± 0.0021
Direct Method*	0.032 ± 0.0001	-
Most Frequent†	-0.023 ± 0.0001	-
Random Policy	-0.023 ± 0.0001	0.478 ± 0.0026
Estimated Inverse Propensity Score	-0.025 ± 0.1009	0.504 ± 0.0071
Estimated Translated Inverse Propensity Score	$\mathbf{-0.143 \pm 0.0099}$	$\mathbf{0.471 \pm 0.0060}$

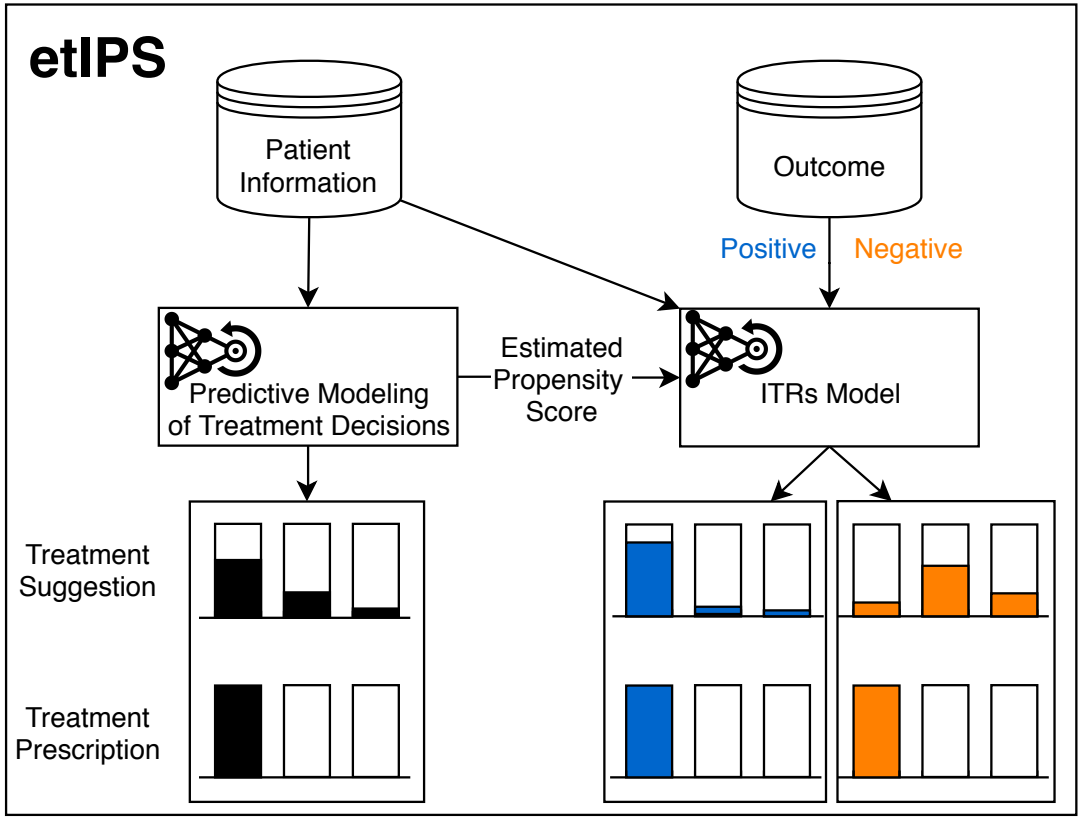
*There is no probability given by $\arg\min_a \mathbb{E}[\delta|X, a]$. The values for DR can therefore not be computed.
†A deterministic policy to suggest the most frequent treatment. No probability information is involved.

Komorowski, M., Celi, L. A., Badawi, O., Gordon, A. C., & Faisal, A. A. (2018). The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. Nature medicine, 24(11), 1716-1720.
Zhou, X., Mayer-Hamblett, N., Khan, U., & Kosorok, M. R. (2017). Residual weighted learning for estimating individualized treatment rules. Journal of the American Statistical Association, 112(517), 169-187.
Dudík, M., Langford, J., & Li, L. (2011, June). Doubly robust policy evaluation and learning. In Proceedings of the 28th International Conference on International Conference on Machine Learning (pp. 1097-1104).

Summary and Outlook

- A flexible framework to learn optimal ITRs
- Batch learning setup in healthcare domain
- Propensity score estimation
  - Predictive modelling of treatment decisions
- ITRs learning
  - Batch learning from bandit feedback (BLBF) algorithms

- Integration with potential outcome models
- Interpretability / Explanability of the learnt ITRs



Thank you!

Dr. rer. nat. Zhiliang Wu
Research Scientist

Siemens AG
Otto-Hahn-Ring 6
81739 München

E-mail zhiliang.wu@siemens.com

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

SIEMENS