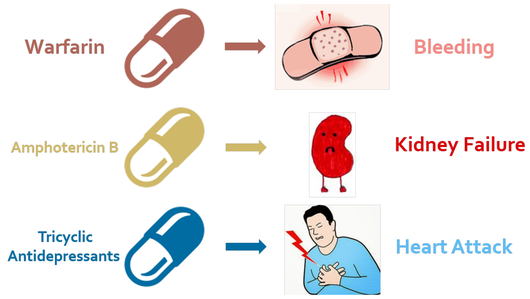
Hawkes Process Modeling of Adverse Drug Reactions with Longitudinal Observational Data

Yujia Bao[1], Charles Kwong[1], Peggy Peissig[2], David Page[1], Rebecca Willett[1]

University of Wisconsin-Madison[1], Marshfield Clinic Research Foundation[2]

Adverse Drug Reaction Discovery

Adverse drug reaction (ADR) discovery is the task of identifying unexpected and negative events caused by pharmaceutical products.



Multiple Self-Controlled Case Series [1]

- Given the time-at-risk window L, for patient p, let

  $x_{p,t,d} \coloneqq$ whether drug d was prescribed at time t,

  $x_{p,t,o} \coloneqq$ whether outcome o was observed at time t.
- Define

$$\tilde{x}_{p,t,d} = \begin{cases} 1, & x_{p,s,d} > 0 \text{ for } s \in \{t-L, \dots, t-1, t\}, \\ 0, & \text{otherwise,} \end{cases}$$

  as the data with imputed missing elements.
- Model the observation using a Poisson distribution

$$x_{p,t,o} \sim \text{Poisson}(\lambda_{p,t,o}).$$

- Parametrize the log-rate of outcome o for patient p at time t as

$$\log \lambda_{p,t,o} = b_{p,o} + \sum_{d \in \mathcal{D}} w_{o,d} \tilde{x}_{p,t,d},$$

  where $b_{p,o}$ is an individual-specific baseline rate and the weight $w_{o,d}$ indicates how predictive drug d is of outcome o.
- Use convex optimization to learn the parameters.
- Limitations: Assume all drugs share the same time-at-risk window and assume no time-varying drug effect.

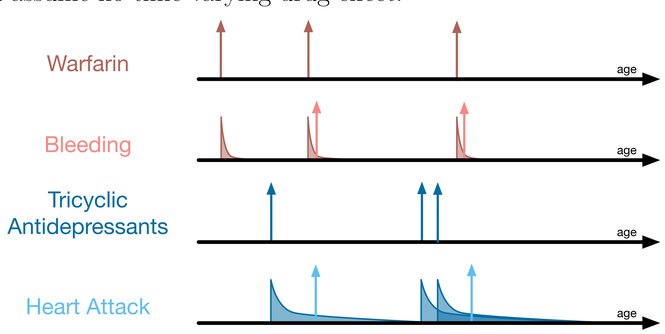


Figure: Visualization of one patient's electronic health record

Hawkes Process

- Hawkes Process is a point process model in which past events influence the likelihood of future events.
- Idea: For each drug-outcome pair, approximate the time-varying effect from the drug to the outcome by a weighted sum of some influence functions ϕ_k.

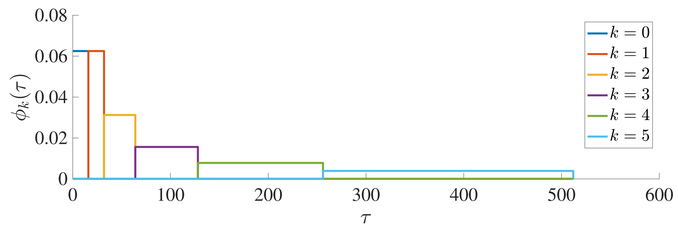


Figure: Piecewise constant influence functions that we used in the experiments. Each ϕ_k gives a normalized count of how many events occurred in some time interval in the past.

- Let N_p be the total number of events observed for patient p. Describe his/her i-th event by the time $\tau_{p,i}$ and the type $m_{p,i}$.
- Model the log-rate of the Hawkes process as following:

$$\log \lambda_{p,o}(\tau) = b_{p,o} + \sum_{\substack{i \le N_p: \\ \tau_{p,i} < \tau \\ m_{p,i} = d}} w_{o,d,k}\, \phi_k(\tau - \tau_{p,i}).$$

The weight $w_{o,d,k}$ indicates how well we may predict outcome o based on a patient being on drug d according to the k-th influence function.

- The log-likelihood for patient p's occurrences of outcome o:

$$\log \ell_{p,o}(b_{p,o}, \boldsymbol{w}) = \sum_{\substack{i \le N_p: \\ m_{p,i} = o}} \log \lambda_{p,o}(\tau_{p,i}) - \int_{\tau_{p,1}}^{\tau_{p,N_p}} \lambda_{p,o}(\tau)\, \mathrm{d}\tau.$$

Regularized maximum likelihood estimator:

$$(\boldsymbol{w}, \boldsymbol{b}) = \arg\min_{\boldsymbol{w}, \boldsymbol{b}} - \sum_{p=1}^{P} \sum_{o \in \mathcal{O}} \log \ell_{p,o}(b_{p,o}, \boldsymbol{w}) + \lambda \|\boldsymbol{w}\|_1.$$

Solve this convex problem by coordinate descent + FISTA.

Dataset

We employ a de-identified version of Marshfield Clinic health system's Electronic Health Records. We extracted 10 drug prescription records and 10 diagnosis records based on the definition of OMOP.

Table: Summary statistics of the cohort

# patients	327,824
# adverse health outcomes	1,940,681
# drug prescription records	11,211,769
# avg. observation duration	9.1 years

Results

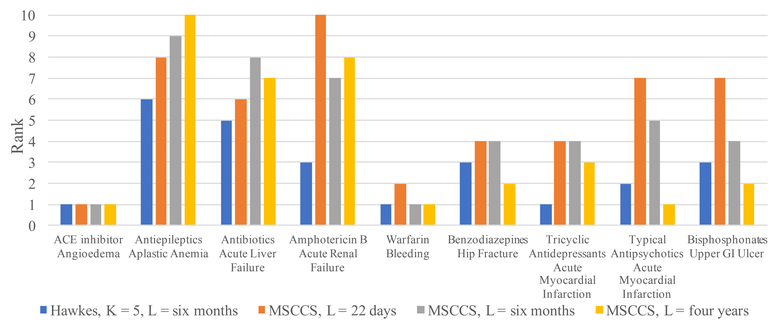


Figure: Rank of true ADR-causing drug among all ten drugs for each true ADR pair

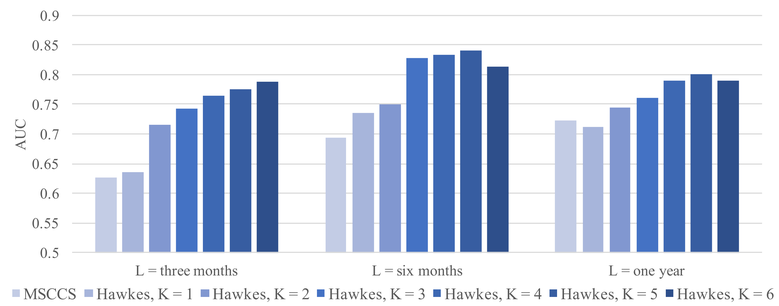


Figure: AUC for MSCCS and Hawkes with various L and K

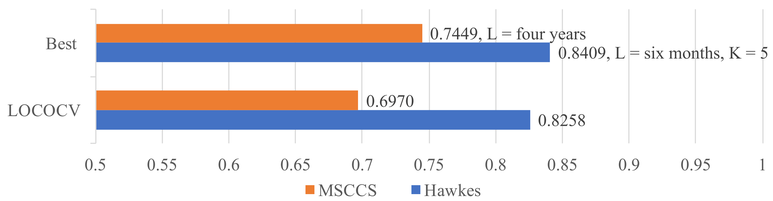


Figure: AUC of the best performers and leave-one-condition-out cross validation.

Acknowledgment

The project described was supported by the NIH BD2K Initiative grant U54 AI117924, the NIGMS grant 2RO1 GM097618, and the Clinical and Translational Science Award (CTSA) program, through the NIH National Center for Advancing Translational Sciences (NCATS), grant UL1TR000427, and by the NSF grant CCF-1418976. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or NSF.

References

[1] Shawn E Simpson, David Madigan, Ivan Zorych, Martijn J Schuemie, Patrick B Ryan, and Marc A Suchard. Multiple self-controlled case series for large-scale longitudinal observational databases. *Biometrics*, 69(4):893–902, 2013.