TL-Lite: Temporal Visualization and Learning for Clinical Forecasting

Jeremy C. Weiss JEREMYWEISS@CMU.EDU
Heinz College of Information Systems and Public Policy
Carnegie Mellon University
Pittsburgh, PA, USA

Editors: Emily Alsentzer[⊗], Matthew B. A. McDermott[⊗], Fabian Falck, Suproteem K. Sarkar, Subhrajit Roy[‡], Stephanie L. Hyland[‡]

Abstract

Clinical data extraction is a necessary step for quantitative analysis in clinical research. Whereas most machine learning algorithms learn from fixed length or regularly-collected panel data, health records data are neither. To facilitate the development of transparent and reproducible machine learning models from such data, we introduce TL-Lite, a clinical data ingestion, transformation, and visualization tool for conducting temporal machine learning. The central principle behind TL-Lite is to provide visual responsiveness at the individual level alongside management of the desired transformations behind the scenes that go on to be applied throughout the cohort and that result in cohort-level summaries, statistics, models and predictions. Characterization of the tool, discussion of design choices, and examples of use demonstrate its added value. A demo is provided at https://www.andrew.cmu.edu/user/jweiss2/viz.html.

Keywords: Clinical forecasting, visualization, extraction, survival analysis

1. Introduction

Time is a focal dimension of clinical data, and visualizations used in clinical practice center around it. For example, discussions of patients in clinical rounds involve updates on status and are often convened around visual aids that depict measurements of progression and recovery. For example, public health intervention planning is benefitted from knowing prediction lead times, and visualizations that include time lag establish trust in the feasibility of the approach. For example, coordinated care protocols are complicated, and measurement of protocol adherence can be monitored using visualizations with temporal landmarks of receipt and monitoring for adverse events. These examples of healthcare delivery all show the importance of time. That predictive forecasting can guide anticipatory decision making is a strong argument for the role of temporal machine learning in health care.

Since the widespread adoption of electronic health records, significant advances have been made in clinical data visualization and clinical forecasting. Yet there remains a gap between these efforts. While it is standard to provide visual evidence of the value of a model for any model developed, its further applied use is frequently limited. One source comes from the lack of transparency in the end-to-end computation, potentially due to a laborious, error-prone, or opaque transformation of the clinical data stream. As an example, the standard and pervasive

Cox model leaves feature timing as implicit, and only inferentially are the features associated with time $t = 0$. How multiple measurements across time are reduced to single value representations may be left to text or to imputation. On the other hand, end-to-end prediction systems must address the problem that the user design choices are specified up front. Often making good choices is contingent on intermediate computations realized through checkpoint visualizations. The importance of these checkpoints often stems the clinical utility of end-to-end designs.

To address this gap, we present TL-Lite, a reactive and visual end-to-end system for conducting clinical forecasting. TL-Lite begins with visualizations of database extracts in the form of an entity-attribute-value (EAV) model, and ends with visual assessments of an internally validated, user-specified, temporal model. Along the way, users can see the effects of their design choices through visual summaries at the individual and cohort level. This enables them to better understand their data and to refine and adjust settings for their analysis. It also facilitates comparative assessments of model stability, feature irreplaceability, and choice of phenotype definition.

While the individual elements of TL-Lite are well known, their integration into an interactive clinical research tool is novel and useful. In fact, possessing similarities to standard tools and following established guidelines serves to enhance familiarity and increase ease of use. Here are two similarities. Like leading commercial and clinical research visualization tools, time is given a visual dimension, and the data is shown at the level of the individual. And like these systems, users can navigate patient data across views to provide high-level characterization and to inspect low-level data elements.

Unlike these systems, a user can design a forecasting model and assess its performance in minutes. We are unaware of any single visual framework with individual-level and cohort visualizations, user-specified model construction, and forecasting assessments. The closest systems in spirit are those embedded in commercial EHR systems, which have been described in validation efforts (Bennett et al., 2019). However, the degree of manual implementation appears high, and the time-to-analysis of these systems appears slower by orders of magnitude. In terms of time-to-analysis, the Green Button informatics consult service reports a turnaround time of 1 to 3 days. However, this system appears to be split into separate tools connected via experienced personnel (Longhurst et al., 2014; Schuler et al., 2018).

Contributions. We introduce Temporal Learning Lite (TL-Lite), an interactive visualization tool for constructing temporal forecasting models from clinical entity-attribute-value data. We motivate and describe design choices that integrate visualization and machine learning elements. We demonstrate the tool with three use cases based on two large private data sets and a public subset. A demo of the tool is available at: `https://www.andrew.cmu.edu/user/jweiss2/viz.html`, with credentials of login: demo, and password: omed.

Related work. The ecosystem of clinical visualization and clinical machine learning tools is rich and comprises too many to name, so we will highlight selected works instead. Please see the recent review which provides a detailed discussion of this space (Levy-Fix et al., 2019). TL-Lite follows the line of research into visualizations with time occupying a visual axis (Plaisant et al., 1996; Shahar et al., 2003; Aigner and Miksch, 2006). Like CareVis, the cohort visualization leverages the concept of view coupling, where multiple plots and tables react together synchronously (Aigner and Miksch, 2006). For clinical event selection, visual fea-

ture selection augments list-based and regular expression-based selection and has been explored previously (Krause et al., 2014). TL-Lite is built using the R Shiny framework and leverages many of the tools in the ecosystem (Chang et al., 2017). Similar to Clairvoyance (Jarrett et al., 2020), TL-Lite is both temporal and end-to-end, with the benefit of simplicity and interactive visualization during interaction with the user. Guidelines meant to assist cohort specification and prognostic modeling are published as checklists (Collins et al., 2015; Norgeot et al., 2020), and they provide effective mental models and are referenced in TL-Lite.

2. Method

End users of the tool, *e.g.*, health professionals and clinical researchers, will be familiar with the underlying data elements. We want them to use the tool without requiring detailed knowledge of the algorithms that organize, process, and model the data. With this in mind, we follow several key concepts in the TL-Lite design.

Time as a visual axis. Whereas many machine learning models treat time as secondary to value by making time implicit, the de facto framework of electronic health record visualization is the timeline. In timelines, data elements of interest, *e.g.*, values, intervals, sequences, are displayed across time as a visual axis. Understanding how raw data came to be substituted with processed values is critical when addressing stability, robustness, and applicability, to name a few desiderata. Observing values over time enables prediction lead time specifications, measurement error identification, and may uncover relationships between events that may impact study relevance and measurement review.

Reactivity and completeness. Alert fatigue, burnout, and technical debt are associated with increased use of electronic health records. We design TL-Lite to be responsive to minimize increased burden and worsening of the user experience. The inherent tradeoff is that reactivity is possible for individual data and light computation, while cohort data is complete and representative, yet may be large and slow to process.

Transparency of representation. While the data stream may be difficult to consume in entirety, the ability to inspect and "sanity-check" intermediate representations establishes trust. By building visualization elements that enable exploration of these representations, the user learns to rely on these elements as the complexity of the modeling effort increases. Checkpointing these representations and using bookmarking to save user design choices aids the conduct and sharing of reproducible research.

Complementarity. The central goal of TL-Lite is to facilitate well-specified and well-crafted predictive forecasting, and the visualization tool is meant to ease this process. At the same time, organizing the clinical data stream into meaningful visualizations can be aided by introducing machine learning elements. There is a complementarity of these approaches, where leveraging the benefits of one where another hits roadblocks results in a better overall solution.

2.1. TL-Lite Overview

TL-Lite is an a R Shiny temporal data processing pipeline (Figure 1). It accepts delimited files which specify at minimum four columns representing: **(Patient ID, Time, Event, Value)** and which are ordered by Patient ID and Time. The application displays data at the individual and cohort level, and it allows custom processing of the patient timelines into matrices and other objects for temporal machine learning. It pro-

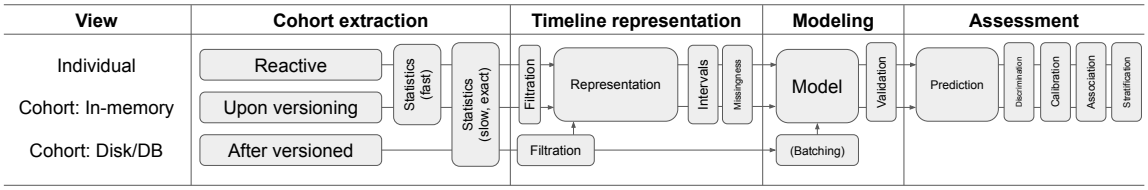

Figure 1: TL-Lite pipeline. Each stage is interactive, enabling the user to make adjustments in each stage. Versioning of intermediate computations is modular, and passing versioned computations from panel to panel enables reuse and refinement.

vides an interface for building models and assessing those models with standard survival performance measures. Intermediate computations can be saved and downloaded and then reuploaded for later use. With training, users can build custom forecasting models to assess risk for any number of cohort, outcome, and feature combinations.

The application is designed to be reactive on local data and at the same time act as an effective representation for the complete cohort. To achieve this, the data is split into three views: (1) the individual, (2) the in-memory cohort, and (3) the disk/database (DB) cohort. Data for an individual are displayed with time as a visual axis for familiarity and as an exemplar for the effects on an individual when selection criteria are applied. The in-memory cohort acts as a representative subsample of the larger out-of-memory cohort and is used to determine the clinical event list and to compute statistical summaries and missingness. The disk/DB cohort can be optionally integrated into these summaries throughout the processing pipeline, and precision comes at the expense of slower calculations. The usability of TL-Lite relies on a presenting reactive visualizations to develop the desired pipeline, which is then followed by a slower set of steps once the pipeline has been determined.

2.2. Panels of TL-Lite

TL-Lite is split into four main panels for (1) cohort extraction, (2) representation, (3) modeling, and (4) assessment. It also has accessory panels containing guideline checklists and fixed width and panel data conversion.

Cohort extraction panel. The initial visualization displays individual-level data in the form of a patient trajectory represented by events indexed by a number (the event index) across time (Figure 2). The user has options to explore the data by switching patients, brushing the data points for inspection, selection, and zooming. Faceted value plots show value changes across time. Cohort definitions are crafted using option groups for the temporal windowing, outcomes, features, inclusion and exclusion criteria, and custom annotations. Temporal windowing specifies the time window in a patient trajectory where a loss function is evaluated. Outcomes and features define which clinical events to maintain and which may be discarded from future analysis, as do exclusion (if any) and inclusion (if any) criteria. Annotations provide a simple way to add events of interest. For example, the outcome "Severe thrombocytopenia" can be the result of an annotation measuring when the platelet count is measured to be below 50×10^9/L. Another example is the definition of a marker for 24 hours post ICU admission which can

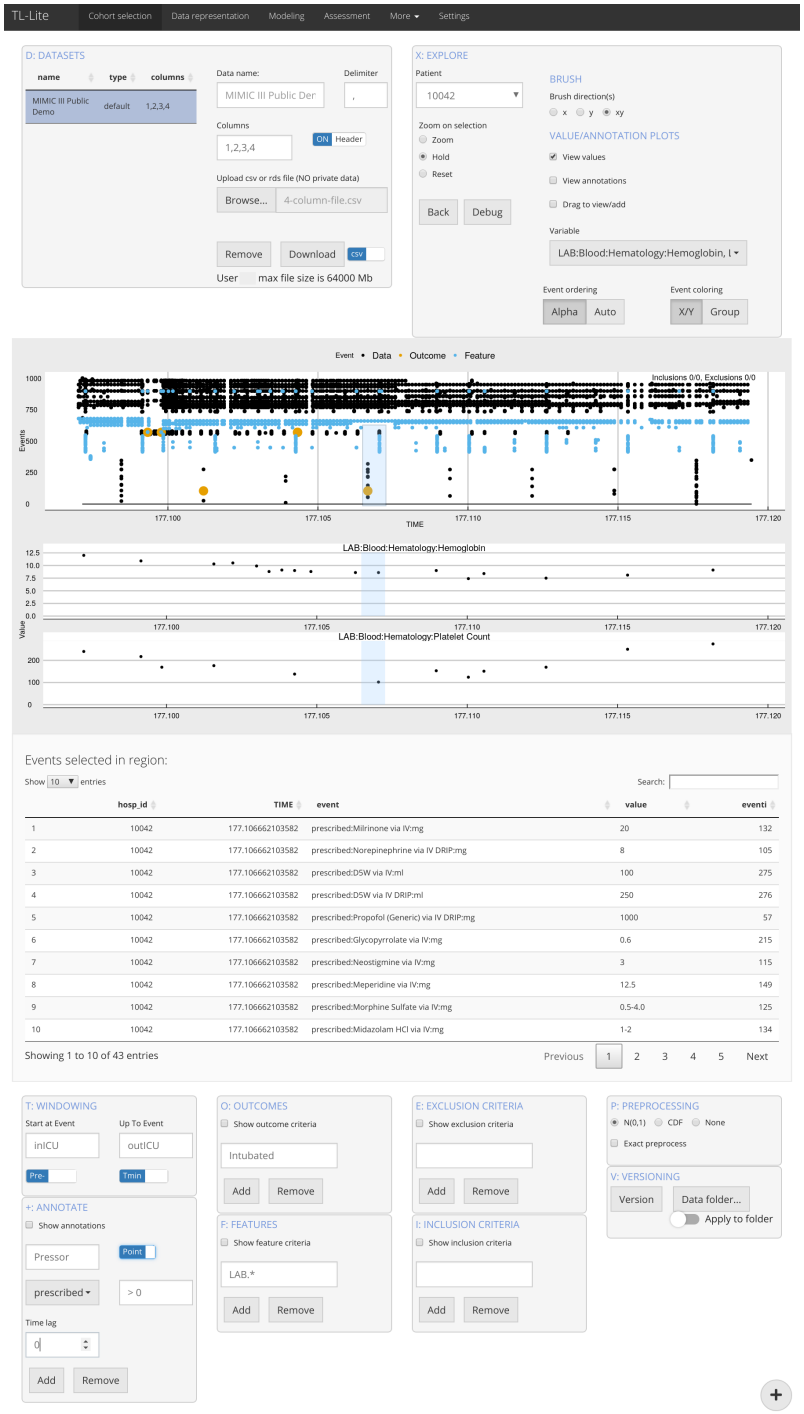

Figure 2: Cohort selection panel. A patient trajectory from the selected dataset is displayed with user-selected value plots. Highlighted events data are shown in the table. Cohort refinements are applied to the patient graph reactively, and upon versioning are applied to the in-memory (and optionally the disk/DB) cohort.

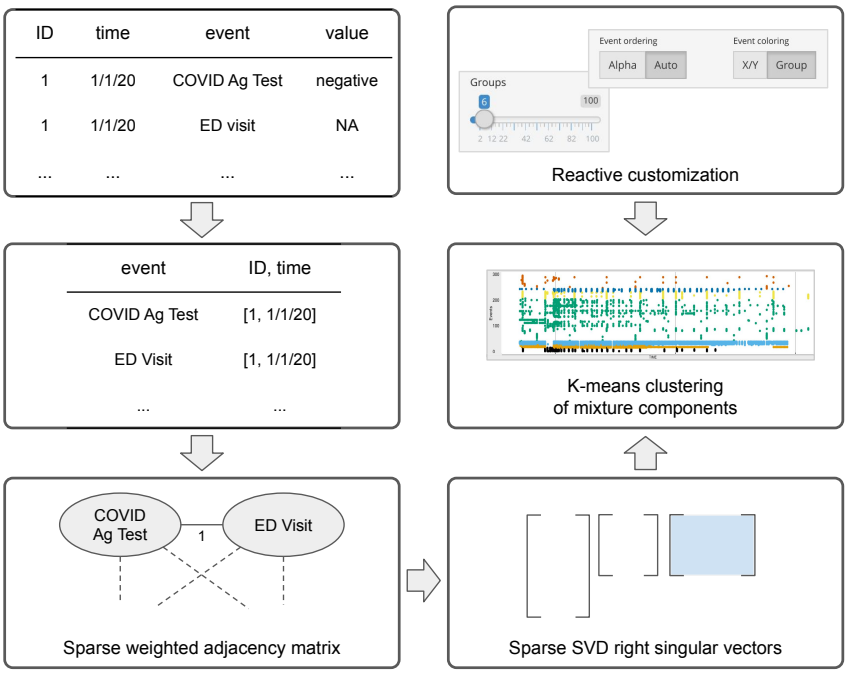

Figure 3: Automated groupings of clinical events (middle right panel, y-axis) eases the process of drag-to-select feature selection. Sparse SVD applied to sparse adjacency matrices based on temporal co-occurrences result in right singular vectors (mixture vectors) that are clustered with k-means.

be used as a lower window bound for analyses that would like to discount outcomes occurring in the first 24 hours after admission.

Event selection can be a cumbersome task at baseline, yet it is a key element in outcome and feature specification, *i.e.*, electronic phenotyping and covariate selection respectively. The user would like to pull some subset from the large number of distinct clinical event types, and selecting from very long lists may be prohibitively slow. We provide several alternatives: list selection, regular expression selection, and brushed selection. Regular expression and brushed selection enable quick assignment of features, while list selection is effective in refining the subsequent selection list.

Brushed selection is affected by the ordering of events because selection is limited to a contiguous interval along a visual axis. Alphabetically-ordered features may be meaningfully presorted if in their name they contain a hierarchy, *e.g.* International Classification of Diseases 10 (ICD-10) and Logical Observation Identifiers Names and Codes (LOINC) codes. However, the user may have events not coded this way or may prefer alternate groupings. To provide additional event groups, we provide an auto-group feature which reactively identifies event clusters (Figure 3) so that contiguous brushed selection selects like events. The clustering is on EAV data, so a conversion to a feature matrix is necessary for standard clustering algorithms. This is accomplished by specifying an event-event weighted adjacency matrix, where the value of pairs corresponds to the number of times the events co-occur in pa-

tient ID and time. This could be a slow and memory intensive process—quadratic in the number of events and linear in the number of event occurrences—so a sparse and approximate approach is chosen to maintain reactivity. Events are subsampled from the selected individual and a subset of individuals in the in-memory cohort, and a sparse weighted adjacency matrix is constructed. A rank-d sparse singular value decomposition (SVD) is applied, and the right singular vectors are interpreted as mixture components. Clustering using k-means identifies events with similar mixtures which correspond to similar co-occurrence profiles. The number of clusters k is user-selected.

When the user is satisfied with the cohort selection criteria, they can version their dataset. This instructs the server to compute the extraction for the cohort with options for application to the in-memory cohort only (faster) or to the disk/DB cohort as well (slower). The versioned objected stores the resulting EAV data as well as design parameters involved in its creation. The versioned object can then be selected for review of the changes applied. Per-feature value processing is available with z-normalization and quantile transformations, and the original values are kept in an accessory column for summary statistics processing later. Upon completion of cohort extraction, the user selects the representation panel and passes the versioned "cohort" object for representational refinement.

Representation panel. Because the extracted data remains in EAV format, a transformation into subsets is necessary for forecasting analyses. The primary division comes through the definition of time $t = 0$, where two representations are computed from EAV data: (1) a baseline representation for events timestamped prior to $t = 0$, and (2) a forecasting representation (including outcomes, time-varying covariates, and windowing parameters) for those timestamped with $t > 0$. While the lower window bound from cohort selection is a natural $t = 0$ definition, the user may select alternatives. The $t > 0$ representation is a set of interval EAV data so that time-varying analyses have the necessary underlying data to compute from. The representation of data and imputation technique (if necessary) is also selected.

Shown to the user are baseline ($t \leq 0$) summaries including event missingness profiles and summary statistics (exemplified in Figure A.2. These are helpful in identifying features that will require high levels of imputation or that possess values that are unexpected. Alongside baseline visualizations, the forecasting representation is depicted in a scope, where the interval determinations are user-selected. In the "forecast" view, time is shown in both the x- and y- axes, where the x-axis indicates the time of prediction and feature availability, and the y-axis indicates the time of forecast, *i.e.* the time of outcomes and loss measurement. For example, the time-invariant Cox model will be represented as a vertical line with $x = t = 0$ because only features at or before baseline may be used in the forecast. In the "trajectory" view, the EAV data for an individual are shown with the interval markers indicating which interval the data will be stored in. When the user is satisfied, the version button is clicked, and a versioned "representation" object is created.

Modeling panel. With the "representation" in hand, the user can now specify the type of machine learning testing framework, with options of "train only", "train/tune/test", and "cross-validation". Selection of an option creates an object based on the patient IDs stored in the representation object, with the user specifying sizing hyperparameters. The outcome type and models are selected, and for now the options

are time-to-event using glmnet Cox and Cox. Hyperparameters for glmnet Cox are chosen by running the *cv.glmnet* function which conducts inner cross-validation determining level of shrinkage, which is separate from the outer cross-validation user option. When the user is satisfied, the validation and representation objects are passed to the selected model routine, and a "modeling" object is produced.

Assessment panel. We follow the TRIPOD guidelines that suggest assessment measures of discrimination, calibration, and feature association (Collins et al., 2015). The user specifies evaluation on the train, tune or test set, with the default being the tune set so as not to over-analyze the test set. For discrimination, we provide the concordance measure with confidence intervals based on bootstrap resampling of the data being evaluated. Three visual assessments are also provided: forest plots, Kaplan-Meier risk stratification plots (Kassambara et al., 2017), and survival calibration plots. Examples are given in Figure 4. The forest plot, which is independent of assessment data, is displayed per validation fold with a single point per feature given by the geometric mean. When the number of strata and prediction times are entered, strata based on predicted survival are created and assessed with the Kaplan Meier group survival estimate. Note that, despite using a Cox model that observes the proportional hazards assumption, survival probabilities must be computed using both the model and a base hazard (we use the Breslow estimator per fold), and so across cross-validation folds the probabilities and rankings are in fact time-dependent. When the user is satisfied, the data and figures can be downloaded for further analysis.

3. Experiments and Results

We demonstrate use of TL-Lite in two electronic health records from MIMIC III and in one from the Marshfield Clinic Johnson et al. (2016, 2019). MIMIC III has a public subset of 100 deceased individuals and serves as data for the demo: `https://www.andrew.cmu.edu/user/jweiss2/viz.html`, credentials u: demo, p: omed. We investigate the following clinical problems: (1) predicting severe thrombocytopenia during ICU stays among patients with sepsis (meeting Sepsis-3 criteria in MIMIC III); (2) predicting survival among patients admitted to the ICU one day after admission (MIMIC III Public Demo); and (3) predicting microvascular complications of type 2 diabetes mellitus (T2DM) among patients with T2DM. We provide details of the first case, and discuss the second and third cases in Appendices A.1 and A.2.

Severe thrombocytopenia, or a platelet level below 50 $\times 10^9$/L, can lead to excessive bleeding, and in some patients with this level of platelet depletion or lower, transfusing platelets as a blood product can be life saving. Compared to other blood products, the shelf-life of platelets is short (no more than 5 days), so blood banks are interested in matching supplies of platelet products to anticipated demand. For example, a model of time to severe thrombocytopenia could be used in forecasting demand among future sepsis populations. Given that predicting severe thrombocytopenia throughout the course of an ICU stay is a challenging task given only information prior to ICU admission, we do not expect high discriminative ability, due to the uncertain nature of disease progression. Despite this, we may be able to accurately assess level of risk through risk stratification and calibration assessment, since such estimates would be sufficient to forecast demands for subsequent encounters.

Table 1: Severe thrombocytopenia

	Concordance c, [2.5%, 97.5%]
Train	0.701 [0.593, 0.800]
Tune	0.692 [0.586, 0.776]
Test	0.689 [0.576, 0.783]

To address this scenario, we use TL-Lite to construct a time-to-event model based on pre-ICU data. In MIMIC III, the majority of clinical data tied to the patient that are time-stamped prior to entry into the ICU are demographic or laboratory measurements, so we choose to focus on these. We identify patients matching Sepsis-3 criteria following Johnson et al. (2018), process the tables to EAV format, sort by Patient ID and Time, and load into TL-Lite. We define time $t = 0$ as entry into the ICU and the upper window boundary as ICU discharge. We select a z-normalized, per-feature representation with last observation carry forward imputation with mean imputation for numeric measurement types never occurring. We apply 5-fold nested cross-validation.

The resulting model demonstrates moderate discriminative ability indicated by concordance levels of 0.689 (Table 1). The predictive model for risk stratification cleanly separates risk groups (Figure 4), where the highest and lowest quintiles by risk have 80% and 97% 10-day survivals (from severe thrombocytopenia). Groupwise, the predictions are well calibrated. The high risk calibration group has a prediction tail, which can be explored by the user by increasing the number of calibration groups up to sample size limitations (the Greenwood confidence interval widths will also increase). Forest plots and summary statistics are in Figures A.1 and A.2.

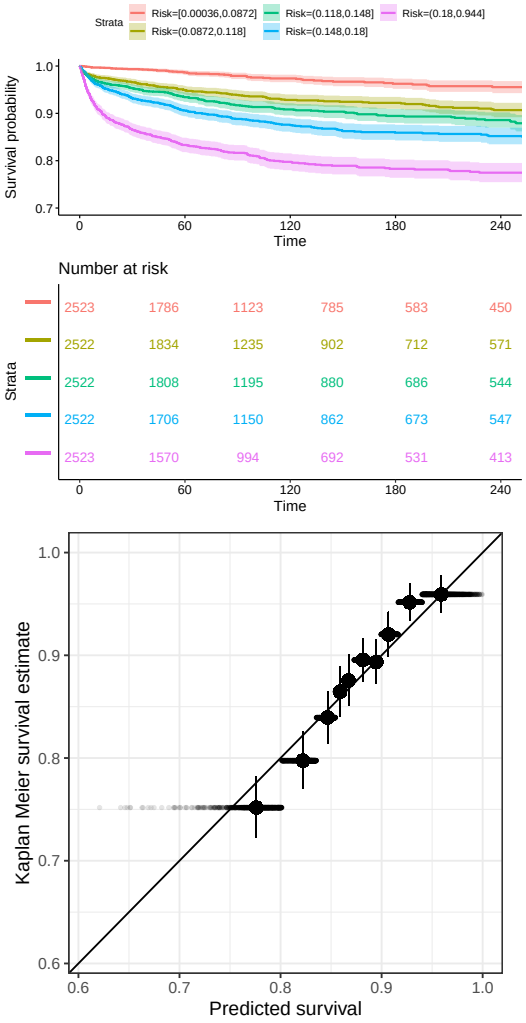

Figure 4: Time (hours) to severe thrombocytopenia ($<50\times10^9$/L), legend: predicted risk intervals. Calibration plot: predicted survival (x-axis) and Kaplan-Meier survival estimates (y-axis).

4. Discussion

TL-Lite is a visualization and forecasting tool to bridge the gap between clinical visualization and machine learning analysis. The tool can assist those with clinical experience in crafting cohorts to address their inquiries,

and it can inform computational researchers about clinical data and highlight the design choices behind processed data formats. For those with experience in both domains, TL-Lite provides ease of use when crafting data sets, baselines for comparison, and visualizations in the cohort design process. We foresee the primary use cases of TL-Lite as twofold: (1) as a useful tool to address clinical equipoise at the point of care, and (2) as an educational tool to bridge the gap between clinical domain experts and machine learners.

As an aid to health professionals. We anticipate in-house use of the tool, where an IT/research team conducts an extraction into EAV data and builds a pipeline for automated updating. Then, approved members can access EAV data on intranet to conduct internal analysis. This use case highlights one of the advantages of TL-Lite, namely that after the initial pipeline setup, end users can be health care professionals, *i.e.*, domain experts, who may not have experience or bandwidth to conduct extensive coding.

As an educational tool. For use as an educational tool, we provide integration with the MIMIC III Public Demo and a panel for converting fixed-width and panel data into EAV form. We will continue to host a demo version, which runs on a cloud micro instance with 1 CPU core and 2 GB memory, and will take requests to host instances with higher and adaptive computational needs. This will aid machine learning for health researchers, who benefit from being able to visualize and interact with electronic health records data to understand their complexity.

On privacy. As with any analysis of health data, privacy is a concern. While credentialing is a service provided in TL-Lite, additional steps should be in place to protect the data. First, the primary intended use case is for intranet deployment, similar to that of many electronic health records. Because of TL-Lite's download functionality, either users must not have access from personal devices (intranet use case) or the data must be deemed fully de-identified (educational use case), or both. One possibility is to only allow viewership and download of the in-memory cohort, which is determined to meet de-identification standards and that satisfies access controls, *e.g.*, is in a provider's or care unit's active patient set. Then, versioning actions could be applied to the disk/DB cohort, with only aggregate statistics presented. This approach could be paired with additional security measures and differentially private algorithms.

On simplicity versus flexibility. For any analytic tool or software, there is a simplicity-flexibility tradeoff. For maximum flexibility, one could omit the entire TL-Lite interface. For users with needs in this direction, TL-Lite provides downloadable extracts at each step, which can be loaded externally for further analysis. This flexibility comes at the cost of not providing or slowly constructing visual aids that are helpful in the analytic process. Alternatively, being restricted to the user interface limits the analytic design choices to those implemented into the system, but does allow for reactive analysis. The tension of flexibility and security (as above) will likely split the health professional aid use case into two; one targeting a research environment with greater control, *e.g.*, user-defined code options for expert coders and research teams, and one targeting clinical experts who seek efficient simplicity. In either case, providing additional documentation, tutorials, and demonstrations will increase TL-Lite's ease of use.

Limitations and future work. There are limitations to TL-Lite. Users who possess datasets with column name mismatches will need to undertake an extensive annotation effort beyond the intended use of this tool. Meta-learning is not integrated. Anal-

yses of distribution shift are limited to applications of available models to new datasets and subgroups, and therefore the internal validation results should be asserted with the caveat that they may not generalize to out-of-sample distributions. Adapting to data shifts for usability in this important context and making improvements to TL-Lite panels are future work (Appendix A.3).

References

Wolfgang Aigner and Silvia Miksch. CareVis: Integrated visualization of computerized protocols and temporal patient data. *Artificial Intelligence in Medicine*, 37(3): 203–218, 2006. ISSN 09333657. doi: 10.1016/j.artmed.2006.04.002.

Tellen Bennett, Seth Russell, James King, Lisa Schilling, Chan Voong, Nancy Rogers, Bonnie Adrian, Nicholas Bruce, and Debashis Ghosh. Accuracy of the Epic Sepsis Prediction Model in a regional health system. *arXiv preprint arXiv:1902.07276*, 2019.

Winston Chang, Joe Cheng, J Allaire, Yihui Xie, Jonathan McPherson, et al. Shiny: web application framework for R. *R package version*, 1(5), 2017.

Gary S Collins, Johannes B Reitsma, Douglas G Altman, and Karel GM Moons. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD). *Circulation*, 131 (2):211–219, 2015.

Daniel Jarrett, Jinsung Yoon, Ioana Bica, Zhaozhi Qian, Ari Ercole, and Mihaela van der Schaar. AutoML pipeline for medical time series. 2020.

Alistair Johnson, Tom Pollard, and R Mark III. MIMIC-III clinical database demo (version 1.4). *PhysioNet*, 10: C2HM2Q, 2019.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

Alistair EW Johnson, Jerome Aboab, Jesse D Raffa, Tom J Pollard, Rodrigo O Deliberato, Leo A Celi, and David J Stone. A comparative analysis of sepsis identification methods in an electronic database. *Critical care medicine*, 46(4): 494–499, 2018.

Alboukadel Kassambara, Marcin Kosinski, P Biecek, and S Fabian. survminer: Drawing survival curves using 'ggplot2'. *R package version 0.3*, 1, 2017.

Josua Krause, Adam Perer, and Enrico Bertini. Infuse: interactive feature selection for predictive modeling of high dimensional data. *IEEE transactions on visualization and computer graphics*, 20(12): 1614–1623, 2014.

Gal Levy-Fix, Gilad J Kuperman, and Noémie Elhadad. Machine learning and visualization in clinical decision support: Current state and future directions. *arXiv preprint arXiv:1906.02664*, 2019.

Christopher A Longhurst, Robert A Harrington, and Nigam H Shah. A 'green button'for using aggregate patient data at the point of care. *Health affairs*, 33(7):1229–1235, 2014.

Beau Norgeot, Giorgio Quer, Brett K Beaulieu-Jones, Ali Torkamani, Raquel Dias, Milena Gianfrancesco, Rima Arnaout, Isaac S Kohane, Suchi Saria, Eric Topol, et al. Minimum information about

clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nature Medicine*, 26 (9):1320–1324, 2020.

Catherine Plaisant, Brett Milash, Anne Rose, Seth Widoff, and Ben Shneiderman. Lifelines: visualizing personal histories. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 221–227, 1996.

Alejandro Schuler, Alison Callahan, Kenneth Jung, and Nigam H Shah. Performing an informatics consult: methods and challenges. *Journal of the American College of Radiology*, 15(3):563–568, 2018.

Yuval Shahar, Dina Goren-Bar, Maya Galperin, David Boaz, and Gil Tahan. KNAVE-II: A distributed architecture for interactive visualization and intelligent exploration of time-oriented clinical data. *Proceedings of Intelligent Data Analysis in Medicine and Pharmacology*, 2003.

Appendix A. Additional Studies

The main paper describes results from the MIMIC III Sepsis-3 cohort in full. The forest plot from the main analysis is provided (Figure A.1). Summary statistics for the cohort characterize these individuals (Figure A.2).

In the process of designing a study for a large cohort, it can be beneficial to map out the process using a subset of the data. TL-Lite facilitates this using reactive interactions with the in-memory cohort which serves as a representative sample of the disk/DB cohort. Here we present statistics that are generated in less than a second (as opposed to seconds to minutes). The advantage of the representative sample approach is that we can compare design choices quickly. For example, the choice of time t=0 is central to survival models, and the resulting data set may be substantially different based on this choice. As an illustration, if we select entry and exit to the critical care unit as two t=0 times, the summary statistics and missingness change substantially (Figures A.3 and A.4). Similarly, a comparison of Figures A.2 and A.3 (left) shows the representativeness of the in-memory cohort with respect to the statistics of the disk/DB cohort.

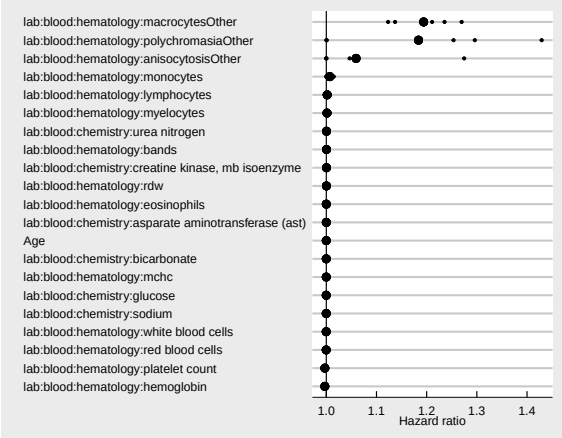

Figure A.1: Thrombocytopenia forest plot

Characteristic	N = 12,612[1]	Percent missing
Age	69 (55, 80)	0
ethnicity:		3.8
white	9,050 (75%)	
black/african american	1,065 (8.8%)	
unknown/not specified	965 (7.9%)	
hispanic or latino	295 (2.4%)	
other	276 (2.3%)	
unable to obtain	254 (2.1%)	
asian	190 (1.6%)	
Other	44 (0.4%)	
gender:		0
m	6,733 (53%)	
f	5,879 (47%)	
lab:blood:chemistry:anion gap	16.0 (13.0, 19.0)	17
lab:blood:chemistry:creatinine	1.20 (0.90, 2.00)	13
lab:blood:chemistry:glucose	129 (104, 171)	18
lab:blood:chemistry:potassium	4.20 (3.80, 4.70)	17
lab:blood:chemistry:sodium	138.0 (135.0, 141.0)	17
lab:blood:hematology:eosinophils	0.40 (0.00, 1.30)	26
lab:blood:hematology:hemoglobin	11.10 (9.60, 12.80)	12
lab:blood:hematology:mcv	90 (86, 95)	12
lab:blood:hematology:neutrophils	81 (71, 88)	26
lab:blood:hematology:platelet count	224 (152, 308)	12
lab:blood:hematology:rdw	14.90 (13.80, 16.60)	12
lab:blood:hematology:white blood cells	11 (8, 16)	12
lab:blood:hematology:pt	14.1 (12.9, 16.7)	20
lab:blood:hematology:ptt	30 (26, 37)	21
lab:blood:hematology:bands	1 (0, 7)	58
lab:blood:chemistry:troponin i	1 (0, 5)	97

[1] Statistics presented: median (IQR) ; n (%)

Figure A.2: MIMIC III Sepsis-3 cohort statistics as reported in TL-Lite

Characteristic	N = 127[1]	Percent missing
Age	68 (56, 80)	0
ethnicity:		0
white	90 (71%)	
black/african american	16 (13%)	
other	7 (5.5%)	
unknown/not specified	5 (3.9%)	
unable to obtain	3 (2.4%)	
asian	2 (1.6%)	
hispanic or latino	2 (1.6%)	
Other	2 (1.6%)	
gender:		0
m	67 (53%)	
f	60 (47%)	
lab:blood:chemistry:anion gap	16.0 (13.0, 18.0)	15
lab:blood:chemistry:creatinine	1.20 (0.80, 1.95)	9.4
lab:blood:chemistry:glucose	140 (106, 185)	15
lab:blood:chemistry:potassium	4.20 (3.77, 4.80)	15
lab:blood:chemistry:sodium	140 (137, 142)	15
lab:blood:hematology:eosinophils	0.40 (0.00, 1.40)	24
lab:blood:hematology:hemoglobin	11.50 (9.80, 13.00)	9.4
lab:blood:hematology:mcv	90.0 (86.0, 94.0)	9.4
lab:blood:hematology:neutrophils	79 (70, 86)	24
lab:blood:hematology:platelet count	225 (169, 288)	9.4
lab:blood:hematology:rdw	14.70 (13.80, 15.70)	9.4
lab:blood:hematology:white blood cells	12 (8, 16)	9.4
lab:blood:hematology:pt	13.9 (13.0, 15.2)	18
lab:blood:hematology:ptt	29 (26, 36)	18
lab:blood:hematology:bands	1 (0, 11)	57
lab:blood:chemistry:troponin i	2.6 (0.8, 7.4)	97

[1] Statistics presented: median (IQR) ; n (%)

Characteristic	N = 127[1]	Percent missing
Age	68 (56, 80)	0
ethnicity:		0
white	90 (71%)	
black/african american	16 (13%)	
other	7 (5.5%)	
unknown/not specified	5 (3.9%)	
unable to obtain	3 (2.4%)	
asian	2 (1.6%)	
hispanic or latino	2 (1.6%)	
Other	2 (1.6%)	
gender:		0
m	67 (53%)	
f	60 (47%)	
lab:blood:chemistry:anion gap	12.0 (10.0, 15.0)	0
lab:blood:chemistry:creatinine	1.00 (0.70, 1.50)	0
lab:blood:chemistry:glucose	118 (97, 155)	0.8
lab:blood:chemistry:potassium	3.90 (3.60, 4.30)	0
lab:blood:chemistry:sodium	141.0 (137.5, 143.5)	0
lab:blood:hematology:eosinophils	0.70 (0.00, 2.00)	7.9
lab:blood:hematology:hemoglobin	10.10 (9.15, 11.30)	0
lab:blood:hematology:mcv	89.0 (86.0, 93.0)	0
lab:blood:hematology:neutrophils	80 (72, 86)	7.9
lab:blood:hematology:platelet count	228 (155, 326)	0
lab:blood:hematology:rdw	15.10 (14.40, 16.55)	0
lab:blood:hematology:white blood cells	10.9 (8.2, 15.6)	0
lab:blood:hematology:pt	13.8 (12.9, 15.4)	4.7
lab:blood:hematology:ptt	31 (27, 39)	4.7
lab:blood:hematology:bands	1 (0, 6)	39
lab:blood:chemistry:troponin i	2.8 (0.9, 5.7)	95

[1] Statistics presented: median (IQR) ; n (%)

Figure A.3: MIMIC III Sepsis-3 in-memory cohort, pre-ICU (left) and post-ICU (right) summary statistics

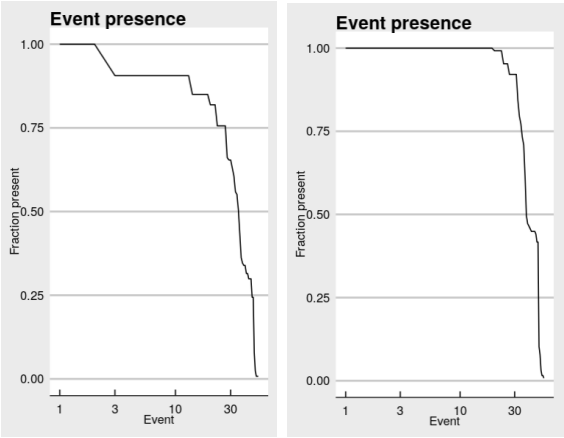

Figure A.4: MIMIC III Sepsis-3 in-memory cohort, pre-ICU (left) and post-ICU (right) missingness statistics, ordered by degree of missingness

In Subsections A.1 and A.2, we show the analysis of multiple additional clinical trajectory datasets to demonstrate the flexibility of TL-Lite in analyzing data with different characteristics and scales.

A.1. MIMIC III Public Demo: Mortality

To facilitate access to rich clinical data with fewer barriers, Physionet.org provides data for 100 deceased individuals in the MIMIC III Public Demo. We build a mortality survival model based on demographics and laboratory results up to the first day after admission to the ICU. Note that inferences from this dataset may be limited due to the knowledge that all patients have passed, which may bias the survival analysis, and due to the limited data size. Despite this, the dataset is effective for visualizing critical care data, and provides an interactive experience particularly useful for those not interacting with electronic health records on a daily basis.

Risk strata and calibration profiles illustrate the limitations of small sample sizes and the difficulty of this forecasting problem: even with two risk strata, the Kaplan Meier group survival confidence bands are overlapping in both plots (Figure A.5). Given these results, interpretations of the forest plot should be considered with caution or withheld entirely.

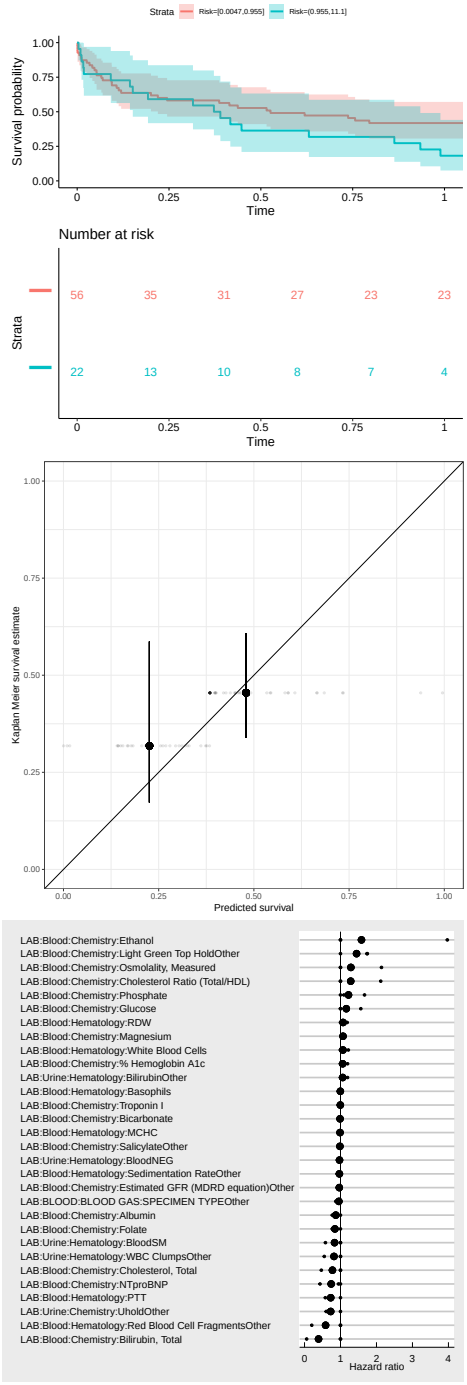

Figure A.5: MIMIC III Public Demo: survival time (years) by risk group, calibration, and hazard ratio forest plot.

A.2. Marshfield Clinic: Complications of Type 2 Diabetes

Marshfield Clinic has a long-standing electronic health record of members in the United States Midwest Region with data from 1960 to the present. We selected a subpopulation diagnosed with type 2 diabetes mellitus to forecast time-to-complication during the period 2007-2017. The health records contain outpatient encounters, diagnoses, laboratory tests, procedures, demographics and medications. We tracked members with new onset T2DM (as determined by an electronic phenotype of ICD code for diabetes, a hemoglobin A1c exceeding 6.4, or an outpatient non-fasting glucose measurement exceeding 200) through this period, with early censorship at the time of last encounter if no encounters were recorded over in 30 months (2.5 years). Outcomes of interest were the first of any microvascular complication including neuropathy (neurologic), retinopathy (ophthalmic), and nephropathy (renal). The motivation is that high risk individuals could be selected for additional elective screening in a public health or wellness program.

Summary statistics describe a characteristic demographic of a Midwest population with elevated hemoglobin A1c levels (Figure A.6). Results indicate the ability to stratify risk and show well calibrated predictions groupwise (Figure A.7). The risk burden is substantial: roughly one third of the highest quintile will experience a complication over 10 years, and even one fifth of the low risk quintile will as well. The forest plot in Figure A.8 suggests that for many individuals the disease has already manifested itself in the form of diagnosis codes indicating diabetes with extra manifestation, but not to the point of using the codes specifically for the downstream microvascular complications. The degree to which this represents variation in coding as opposed to progression of disease warrants further study.

Characteristic	N = 18 207[1]	Percent missing
Age	62 (51, 73)	0
Person|GENDER_CONCEPT_ID		0
MALE|	9 143 (50%)	
FEMALE|	9 064 (50%)	
Person|RACE_CONCEPT_ID		0.4
White|	15 055 (83%)	
Unknown|	2 707 (15%)	
Asian|	189 (1.0%)	
American Indian or Alaska Native|	186 (1.0%)	
Measurement|Creatinine serum/plasma|2160-0		17
Normal	13 048 (86%)	
High	1 948 (13%)	
Other	133 (0.9%)	
Measurement|Cholesterol [Mass/volume] in Serum or Plasma|2093-3		15
Normal	10 485 (68%)	
High	4 821 (31%)	
Other	220 (1.4%)	
Measurement|Hemoglobin|718-7		18
Normal	12 532 (84%)	
Low	1 786 (12%)	
High	619 (4.1%)	
Other	27 (0.2%)	
Measurement|White Blood cell (WBC) count (leukocyte)|26464-8		20
Normal	12 669 (87%)	
High	1 628 (11%)	
Low	291 (2.0%)	
Other	41 (0.3%)	
Diagnosis|Tobacco use disorder|305.1		
Measurement|Hemoglobin A1c in Blood|55454-3		43
High	7 220 (69%)	
Normal	3 190 (31%)	
Other	14 (0.1%)	
Measurement|BP diastolic|	77 (70, 84)	15
Measurement|BP systolic|	130 (120, 140)	15
Measurement|Body mass index|	34 (29, 39)	23
Measurement|Microalbumin/Creatinine [Mass Ratio] in Urine|14959-1		91
Normal	1 380 (84%)	
High	265 (16%)	
Abnormally high	1 (<0.1%)	

[1] Statistics presented: median (IQR) ; n (%)

Figure A.6: Marshfield Clinic summary statistics at onset of type 2 diabetes

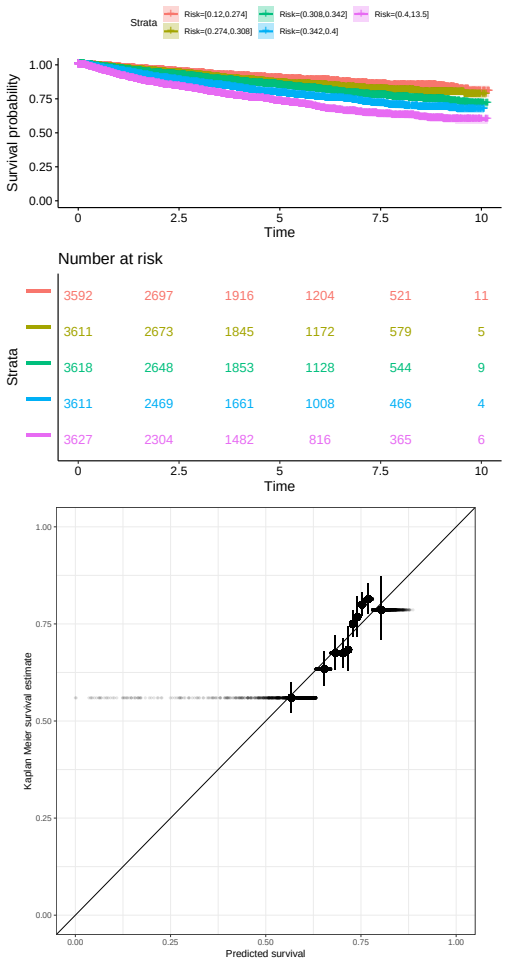

Figure A.7: Marshfield Clinic: survival time from T2DM complications (years), calibration by risk group

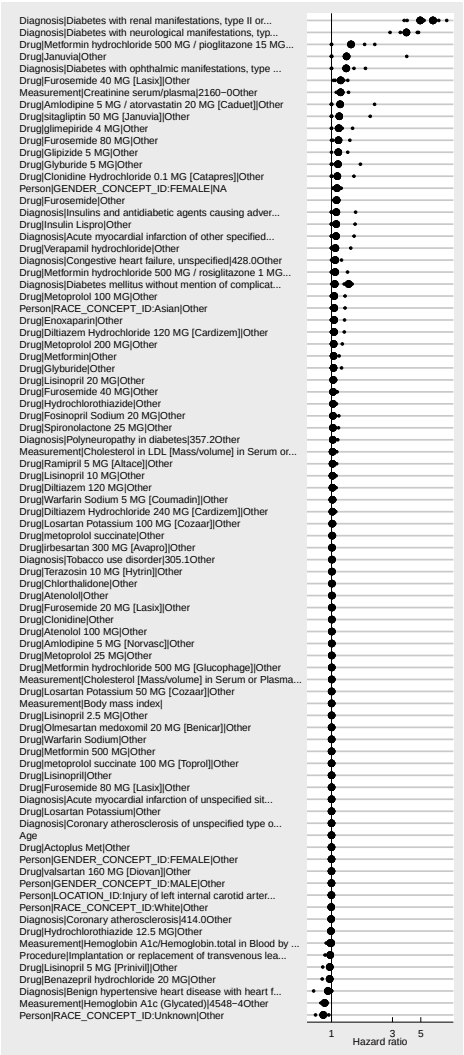

Figure A.8: Marshfield Clinic: hazard ratio forest plot

A.3. Additional Future Work

We maintain a list of requested improvements to the elements of each panel, which includes use of cohort matching and weighting schemes (case-control analysis, inverse weighting, etc.), additional machine learning models, imputation methods, model refinement capabilities, assessment techniques, and event name cleaning tools. The challenge is to maintain low levels of technical debt for simplicity and longevity while providing a sophisticated, interactive forecasting tool. Finally, expansion of the EAV to allow for richer value representation, *e.g.*, intervals, free text, images, and graphs, would enable exploration of other data types while maintaining time as a primary element in the analysis.