Democratizing EHR Analyses with FIDDLE – A Flexible Preprocessing Pipeline for Structured Clinical Data

Shengpu Tang, Parmida Davarmanesh, Yanmeng Song, Danai Koutra, Michael W. Sjoding, and Jenna Wiens

Supplementary Material

1. Additional Details of FIDDLE
   1.1 Assumptions
   1.2 Details of FIDDLE implementation
   1.3 Guidelines for setting FIDDLE arguments

2. Data Extraction
   2.1 MIMIC-III
   2.2 eICU

3. Clinical Outcomes & Study Cohorts

4. Discussion of FIDDLE Execution
   4.1 FIDDLE applied to MIMIC-III
   4.2 FIDDLE applied to eICU

5. Hyperparameters and Model Selection

6. Details of MIMIC-Extract

7. Extended Experimental Results
   7.1 MIMIC-III with discretize = ‘false’
   7.2 Performance of all FIDDLE-based ML models
   7.3 Comparing FIDDLE-based models to additional baselines for MIMIC-III
   7.4 Inspection of learned models
   7.5 Sensitivity Analyses & Flexibility of FIDDLE

8. Applying FIDDLE to Additional Prediction Tasks
   8.1 Seq-to-seq LSTM model
   8.2 Predicting 90-day post-discharge mortality

9. Qualitative Comparisons of FIDDLE to Other EHR Data Pipelines
1. Additional Details of FIDDLE

Here we present important assumptions and guidelines for setting the user-defined arguments. Additional details can be found in the code at https://gitlab.eecs.umich.edu/MLD3/FIDDLE.

1.1 Assumptions

We designed FIDDLE to be largely data-driven, making only a small number of assumptions:

- the unit of \( t \) must be consistent across all rows (e.g., in hours, days, or weeks) and the same for \( T \);
- the variable_value for all rows having the same variable_name are assumed to have the same unit, and must have a consistent time-stamp type (i.e., \( t \) must be either NULL or non-NUL across all rows with the same variable_name);
- the input table should not contain duplicates (i.e., two rows with the same [ID, t, variable_name]).

1.2 Details of FIDDLE implementation

(2) Transform

- Whenever there are multiple recordings within a single time bin, the most recent recording is selected.
- Missing values for “frequent” variables are imputed with carry-forward imputation. As EHR data are often “missing not at random,” the pipeline keeps track of imputed values with a ‘mask’ (indicating presence) and ‘delta time’ (the time since the last real measurement in number of time bins) [1, 2]. We chose carry-forward imputation over other imputation techniques (e.g., mean imputation, multiple imputation, Gaussian processes) because it makes fewer assumptions about the data and can be easily applied in real-time [3, 4].
- When determining the possible categories or the distribution of quintiles, all values (across \( N \) examples, and across \( L \) time bins for time-dependent data) except NULL are used.

(3) Post-filter

- We filter out features that are equal to 1 (or 0) in \( \leq \theta_2 \times 100\% \) of examples. For time-dependent features, the same rule is applied with the additional consideration for time bins: a feature is considered “equal to 1 (or 0)” for an ID if it is always 1 (or 0) over the \( L \) time bins. We then combine each group of pairwise perfectly correlated features (i.e., two features with Pearson’s correlation coefficient of \( \pm 1 \)) into a single feature, saving all original feature names as aliases.

1.3 Guidelines for setting FIDDLE arguments

The user-defined arguments of FIDDLE include: \( T \), \( dt \), \( \theta_1 \), \( \theta_2 \), \( \theta_{freq} \), \( K \) summary statistics functions, and discretization settings. The settings of these arguments could affect the features and how they can be used. We provided reasonable default values in the implementation, and here list some practical considerations: (i) prediction time and frequency, (ii) temporal density of data, and (iii) class balance.

(i) The prediction time and frequency determine the appropriate settings for \( T \) and \( dt \). The risk stratification tasks in our proof-of-concept experiment all involve a single prediction at the end of a fixed prediction window. It is thus most reasonable to set \( T \) to be the length of prediction window. Another possible formulation is to make multiple predictions where each prediction depends on only data from the past (not the future), using models like LSTM or fully convolutional networks. In that case, for example, if a prediction needs to be made every 4 hours over a 48-hour period, then \( T \) should be 48 hours, whereas \( dt \) should be at most 4 hours.
(ii) The temporal density of data, that is, how often the variables are typically measured, also affects the setting of $dt$. This can be achieved by plotting a histogram of recording frequency. In our case, we observed that the maximum hourly frequency is $\sim 1.2$ times, which suggests $dt$ should not be smaller than 1 hour. While most variables are recorded on average $<0.1$ time per hour (most of the time not recorded), the 6 vital signs are recorded slightly $>1$ time per hour. Thus, given that in the ICU, vital signs are usually collected once per hour, we set $dt = 1$. This also implies the setting of $\theta_{freq}$ to be 1. Besides determining the value for $dt$ from context (how granular we want to encode the data), we can also sweep the range (if there are sufficient computational resources and time) given the prediction frequency and the temporal density of data.

(iii) We recommend setting $\theta_1 = \theta_2 = \theta$ and be conservative to avoid removing information that could be potentially useful. For binary classification, the rule-of-the-thumb we suggest is to set $\theta$ to be about 1/100 of the minority class. For example, our cohorts consist of $\sim 10\%$ positive cases, so setting $\theta = 0.001$ is appropriate, whereas for a cohort with only 1% positive cases, then $\theta = 0.0001$ is more appropriate. Given sufficient computational resources and time, the value of $\theta$ can also be swept and optimized.

Finally, for the summary statistics functions $\{\phi_j\}_{j=1}^K$, we included by default the most basic summary statistic functions (i.e., minimum, maximum, and mean). If on average, we expect more than one value per time bin, then one can also include higher order statistics such as standard deviation and rate of change.
2. Data Extraction

2.1 MIMIC-III

We extracted data from all structured tables in MIMIC-III (Table 3A in main text) that record time-invariant or time-dependent data for ICU patients. The following tables from the database were used only indirectly or not used at all:

- **D_CPT, D_ICD_DIAGNOSES, D_ICD_PROCEDURES, D_ITEMS, and D_LABITEMS** are dictionary mappings. Only D_ITEMS and D_LABITEMS were used to search for human-interpretable variable names.
- **CALLOUT, SERVICES, and TRANSFERS** were excluded since they define patients’ movement and locations throughout every hospital admission, and their information has been incorporated in ICUSTAYS.
- **DIAGNOSES_ICD** was excluded since there is no associated timestamp.
- **CPTEVENTS, DRGCODES, PROCEDURES_ICD** contain codes for billing purposes. They do not have accurate timestamps (only CPTEVENTS has the date of each record, while the other two tables are not timestamped) and thus were not used for our analysis.
- **CAREGIVERS** was excluded since it only provides information regarding caregivers, not patients.
- **INPUTEVENTS_CV** was not used as it contains information regarding patients in the CareVue system, so it is not for the cohorts of interest.
- **PRESCRIPTIONS** table has a time precision of days, and only contains medications ordered, not necessarily administered for patients. We relied on INPUTEVENTS_MV for drugs administered patients.
- **NOTEEVENTS** was excluded since it contains unstructured data including clinical notes and diagnosis.

The MIMIC-III database uses ITEMIDs to identify various measurements/recordings (e.g., ‘Heart Rate’ has an ITEMID of ‘220045’). We addressed issues in the extracted data by the following procedures. We tried to make as few assumptions as we could.

- We removed data that are explicitly marked as erroneous in the database. These include: rows in INPUTEVENTS_MV with STATUS=‘Rewritten’ or negative amounts administered, rows in OUTPUTEVENTS with ISERROR=‘1’, and rows in CHARTEVENTS with ERROR=‘1’.
- We removed laboratory values from CHARTEVENTS, since they are copied from LABEVENTS for the purpose of being displayed on the patient’s electronic chart. According to the database documentation, if disagreements between measurements occur, LABEVENTS should be taken as the ground truth.
- For each of the following eight variables – HR, RR, Temp, SysBP, DiaBP, SpO2, height and weight – in CHARTEVENTS, multiple ITEMIDs are used to describe different measurements. With the help of a critical care physician, we remapped and combined each ITEMID group to a single variable. Units of measurements were converted to SI units if they are inconsistent (e.g., C/F for temperature, lb/kg for weight, inch/cm for height).
- In INPUTEVENTS_MV, we converted units, including mass, volume, dosage, and time, to make sure the same ITEMID is always recorded in the same unit. We assumed that any fluid administered to a patient was administered with a constant rate throughout administration. Substances that are administered as a bolus have a duration of 1 minute, and we used the total amount as the value. For drugs infused continually over a period of time, we used the average rate. Amount, rate, and dose were used to describe each ITEMID.
- In DATETIMEEVENTS, the values recorded are timestamps. We converted them to numerical values representing the time elapsed (in fractional hours) relative to the respective ICU admission time (t=0).
• *INPUTEVENTS_MV* and *PROCEDUREEVENT_MV* have STARTTIME and ENDTIME instead of CHARTTIME. We converted records that are continuous events into discrete events by repeating each record multiple times across the time range with different timestamps. The resampling was done based on some predefined time granularity / resampling frequency.

• Timestamp issues:
  - When interpreting the data in *MICROBIOLOGYEVENTS*, we only considered the type of culture that was drawn. We ignored culture results or organism susceptibility results, since timestamps were unavailable and significant delays are possible between specimen collection and culture results (1–2 days). We included only records having collection time with a time resolution to hours.
  - In *LABEVENTS*, the timestamp is the time of fluid acquisition, rather than the time that the test results were made available to clinical staff. The test result usually comes back in a variable amount of time (4-12h), however timestamped results were not available from the database. In this case, we used the collection timestamp to filter the records, with the caveat that the result is only available after the prediction time, and only then a prediction can be made.

Lastly, we converted all data into a uniform format with four columns: [ID, t, variable_name, variable_value]. In general, the ITEMID is used as the variable_name and the recorded value as variable_value. Note that t is the time in fractional hours relative to the ICU admission time. For records with multiple value columns (e.g., a drug has amount of ‘5’mg and route of ‘Oral’), these are mapped to two separate rows, in addition to a 0/1 variable indicating that the drug was administered.

### 2.2 eICU

Given the large sample size, for the eICU data, we made as few assumptions as possible and considered all records with timestamps (in addition to data recorded at admission). In contrast to our work with MIMIC-III, in which we aggregated common vitals sign measurements, we did not perform any manual variable aggregation, since there were too many different possible aggregations to consider. We extracted data from all structured tables in eICU (*Table 3B* in main text) that record time-invariant or time-dependent data for ICU patients. The following tables were excluded:

- *apacheApsVar, apachePatientResults, apachePredVar* pertain to APACHE, a severity of illness score calculated by hand by a research nurse, and thus is typically not available in real time settings.
- *carePlanCareProvider, carePlanEOL, carePlanGeneral, carePlanGoal, carePlanInfectiousDisease* are either rarely used, or contain information related to end-of-life planning, which could lead to label leakage and thus were excluded.
- *hospital* contains details of hospitals covered by the eICU telehealth program, from which this database is built. These data do not directly pertain to patient health; the hospital/ward ID information is recorded in *patient* which is separately extracted.

Due to the large size of this database, we made minimal assumptions when extracting the data (instead of using extensive manual curation). As a result, some of our choices may not be appropriate in other settings. E.g., depending on your problem type you may want to include details regarding end-of-life planning, or use this information to help define your study cohort. We formatted the data into a table with four columns: [ID, t, variable_name, variable_value], where ID is ‘patientUnitStayID’, corresponding to an ICU stay. The timestamp t. for time-dependent data are present in all tables (except *patient*) and are in the form ‘_____offset’. *Table S2.1* summarizes how variable_name/variable_value columns are extracted from each table.
Table S2.1: Data extraction from eICU database.

| Table               | Format                                                                 |
|---------------------|------------------------------------------------------------------------|
| vitalPeriodic       | Column names are used as variable_name; entries are used as variable_value. Only non-missing values are retained. |
| vitalAperiodic      |                                                                         |
| respiratoryCare     | Column names are used as variable_name; entries are used as variable_value. The following columns are excluded: 'currenthistoryseqnum', 'ventstartoffset', 'ventendoffset', 'priorventstartoffset', 'priorventendoffset'. |
| lab                 | variable_name: labname                                                 |
|                     | variable_value: labresulttext, casted to numbers when appropriate       |
| customLab           | variable_name: labothername                                             |
|                     | variable_value: labothervaluertext, casted to numbers when appropriate   |
| infusionDrug        | variable_name: drugname                                                 |
|                     | variable_value: drugrate                                                |
| microLab            | variable_name: culturesite                                              |
|                     | variable_value: organism                                                |
| note                | variable_name: notetype                                                 |
|                     | variable_value: notetext                                                |
| nurseAssessment     | variable_name: cellattributepath                                        |
| nurseCare           | variable_value: cellattributevalue                                      |
| respiratoryCharting | variable_name: respchartvaluelabel                                      |
|                     | variable_value: respchartvalue                                           |
| nurseCharting       | variable_name: Concatenation of values in                               |
|                     | • nursingchartcelltypecat                                               |
|                     | • nursingchartcelltypevallabel                                          |
|                     | • nursingchartcelltypevalname                                           |
|                     | variable_value: nursingchartvalue                                       |
| medication           | First, rows with unknown drug (missing values for both 'drugname' and 'drughicleqno') are removed. variable_name: Concatenation of the values of {'drugname', 'drughicleqno'}, and then concatenate with one of the attribute columns {'dosage', 'routeadmin', 'frequency'} variable_value: table entries in the corresponding attribute columns |
| intakeOutput        | This table is handled in two parts:                                      |
|                     | 1. variable_name: 'cellpath'                                             |
|                     | variable_value: 'cellvaluertext', casted to numbers when appropriate     |
|                     | 2. Each of the columns 'intaketotal', 'outputtotal', 'dialysis总量', 'nettotal' is converted to a separate variable_name with the entries as variable_value |
| admissionDrug       | variable_value: I                                                       |
| admissionDx         | variable_name: recorded values in table columns                          |
| allergy             | • drugname                                                              |
| diagnosis           | • admitdxpath                                                           |
| pastHistory         | • allergynname                                                          |
| physicalExam        | • diagnosisstring                                                       |
| treatment           | • pasthistorypath                                                      |
|                     | • physicalexampath                                                      |
|                     | • treatmentstring                                                       |
| patient             | Timestamp t is set to NULL to indicate time-invariant data. The following columns are used as variable_name, with the entries as the corresponding variable_value: |
|                     | • gender, age, ethnicity, hospitalid, wardid, admissionheight, hospitaladmittime24, hospitaladmitoffset, hospitaladmitsource, unittype, unitadmittime24, unitadmitsource, unitvisitnumber, unitstaytype, admissionweight, |
|                     | The following columns are excluded because they correspond to information at discharge, which are not available at the time of ICU admission: |
|                     | • dischargeweight, unitdischargeTime24, unitdischargeoffset, unitdischargelocation, unitdischargedstatus, hospitaldischargeyear, hospitaldischargeTime24, hospitaldischargeoffset, hospitaldischargelocation, hospitaldischargedstatus |
3. Clinical Outcomes & Study Cohorts

In our evaluation of FIDDLE, we trained ML models to predict in-hospital mortality, acute respiratory failure, and shock. Interpreting each ICU stay as an example, we developed the following pragmatic outcome definitions, incorporating the clinical experience of a critical care physician (M.W.S.).

- **In-hospital mortality.**
  - **MIMIC-III:** In-hospital mortality was identified by comparing a patient’s time of death (‘DEATHTIME’ in *ADMISSIONS*) and their admission and discharge times of the corresponding hospitalization (‘ADMITTIME’ and ‘DISCHTIME’ in *ADMISSIONS*) [5]. We checked these labels against the ‘DOD’ column in *PATIENTS*.
  - **eICU:** The in-hospital mortality status is explicitly recorded in the *patient* table. Positive cases were identified as ICU visits with a ‘hospitalDischargeStatus’ value of ‘Expired’, and the corresponding ‘hospitalDischargeOffset’ was used to calculate the time of death (relative to ICU admission, t=0).

- **Acute respiratory failure** (ARF) is defined as the need for respiratory support with positive-pressure mechanical ventilation [6, 7].
  - **MIMIC-III:** Positive cases of ARF were identified by either documented receipt of invasive mechanical ventilation (ITEMID=225792) or non-invasive mechanical ventilation (ITEMID=225794) in *PROCEDURESEVENTS_MV*, or the documentation of positive end-expiratory pressure (ITEMID=220339) in *CHARTEVENTS*.
  - **eICU:** Positive cases of ARF were identified by either (i) documentation of ‘ventStartOffset’ in *respiratoryCare*, or (ii) recording of ‘peepLimit’ in *respiratoryCare*, or (iii) recording of ‘PEEP’ or ‘PEEP/CPAP’ in *respiratoryCharting*.

- **Shock** is defined as inadequate perfusion of blood oxygen to organs or tissues [8], characterized by receipt of vasopressor therapy [9].
  - **MIMIC-III:** Positive cases of shock were identified by the receipt of the following vasopressors in *INPUTEVENTS_MV*: norepinephrine (ITEMID=221906), epinephrine (ITEMID=221289), dopamine (ITEMID=221662), vasopressin (ITEMID=222315), and phenylephrine (ITEMID=221749).
  - **eICU:** Positive cases were identified by documentation of vasopressor therapy in *infusionDrug* or *medications*. We considered records of ‘drugName’ containing any of the following keywords: ‘norepinephrine’ (aka ‘levophed’), ‘epinephrine’, ‘dopamine’, ‘vasopressin’, ‘phenylephrine’ (aka ‘neo-synephrine’, ‘neosynephrine’).

For ARF and shock, we defined onset time as the earliest time when the criteria were met. In contrast to previous definitions based on ICD diagnosis codes [5], we focused on clinical data indicating onset of events (e.g., mechanical ventilation and administration of vasopressors), since records of ICD codes do not indicate the time of onset and may correspond poorly to the actual diagnoses [10, 11]. Thus, our clinical-based definitions for these two decompensation tasks more accurately reflect the timing of outcomes.

Based on these three outcomes, we defined five prediction tasks, each with a distinct study cohort (Figure 3 in main text; details in Tables S3.1-S3.4). In all analyses, we excluded neonates and children (age < 18) because their physiology and risk factors differ from adults [5, 12]. We did not attempt to exclude patients with treatment limitations (e.g., those who may be placed on comfort measures), given the difficulty in identifying this status reliably across datasets. While this allows us to compare with previous work [5, 13], it could ultimately make the prediction tasks easier and limit the clinical utility of the learned models. For in-hospital mortality, we used $T = 48$ hours to predict whether the outcome would occur after $T$ following existing work [5]. For ARF and shock, we used both $T = 4$ hours and $T = 12$ hours. Examples (ICU stays) with an event onset time before $T$, or discharges and deaths before $T$, were excluded. For the eICU data, we also excluded examples for which the ground truth labels could not be reliably determined due to lack of sufficient documentation [14]. Specifically, for ARF and shock, we excluded admissions to hospitals without any relevant ventilation records or vasopressor records, respectively.
### Table S3.1: MIMIC-III Study Cohort Characteristics.

| Population                      | mortality, 48h | ARF, 4h | ARF, 12h | shock, 4h | shock, 12h |
|--------------------------------|---------------|---------|----------|-----------|------------|
|                                 | N=8,577       | N=15,873| N=14,174 | N=19,342  | N=17,588   |
| Age, median (IQR)               | 66 (54-78)    | 65 (53-78) | 65 (52-78) | 64 (52-77) | 64 (52-77) |
| Sex: F (%)                      | 3,915 (45.6)  | 7,202 (45.4) | 6,545 (46.2) | 8,603 (44.5) | 7,934 (45.1) |
|                                 |               |         |          |           |            |
| **Race**                        |               |         |          |           |            |
| White                           | 6,248 (72.8)  | 11,623 (73.2) | 10,375 (73.2) | 14,000 (72.4) | 12,683 (72.1) |
| Black                           | 877 (10.2)    | 1,895 (11.9) | 1,760 (12.4) | 2,205 (11.4) | 2,083 (11.8) |
| Hispanic/Latino                 | 331 (3.9)     | 633 (4.0) | 578 (4.1) | 793 (4.1) | 746 (4.2) |
| Asian                           | 227 (2.6)     | 440 (2.8) | 388 (2.7) | 527 (2.7) | 468 (2.7) |
| other                           | 894 (10.4)    | 1,282 (8.1) | 1,073 (7.6) | 1,817 (9.4) | 1,608 (9.1) |
| **ICU type**                    |               |         |          |           |            |
| CCU                             | 1,088 (12.7)  | 2,187 (13.8) | 2,085 (14.7) | 2,307 (11.9) | 2,169 (12.3) |
| CSRU                            | 1,120 (13.1)  | 1,327 (8.4) | 621 (4.4) | 1,836 (9.5) | 1,094 (6.2) |
| MICU                            | 3,747 (43.7)  | 7,819 (49.3) | 7,325 (51.7) | 8,911 (46.1) | 8,430 (47.9) |
| SICU                            | 1,563 (18.2)  | 2,883 (18.2) | 2,658 (18.8) | 3,808 (19.7) | 3,600 (20.5) |
| TSICU                           | 1,059 (12.3)  | 1,657 (10.4) | 1,485 (10.5) | 2,480 (12.8) | 2,295 (13.0) |
| **Outcomes**                    |               |         |          |           |            |
| Primary outcome (%)             | 1,031 (12.0)  | 2,900 (18.3) | 1,368 (9.7) | 2,890 (14.9) | 1,364 (7.8) |
| LOS, median (IQR)               | 86 (63-145)   | 44 (26-75) | 42 (25-70) | 46 (27-85) | 45 (26-79) |

Acronyms: CCU, coronary care unit; CSRU, cardiac surgery recovery unit; MICU, medical intensive care unit; SICU, surgical intensive care unit; TSICU, trauma/surgical intensive care unit.

### Table S3.2: MIMIC-III Data Splitting.

We split each cohort into train/val/test respectively. When training LR and RF models, the train/val splits were combined and used for cross-validation. For in-hospital mortality, we used the same splits as the Harutyunyan et al. benchmark[5] except that we further excluded examples not recorded in Metavision. For ARF and shock, we first assigned each patient in the source population to either of train/val/test partition with 70/15/15 ratio, and then split the study cohorts accordingly. We list the counts of patients or examples with the percentage of positive cases (in parentheses) where appropriate.

| Population                      | Total       | Train      | Val        | Test       |
|--------------------------------|-------------|------------|------------|------------|
| In-hospital mortality, 48h     | 8,577 (12.0)| 6,035 (12.5)| 1,278 (11.3)| 1,264 (10.4)|

**Source population**

|                  |            |            |            |            |
|------------------|------------|------------|------------|------------|
| Patients         | 17,710     | 12,397     | 2,656      | 2,657      |
| ICU stays        | 23,620     | 16,604     | 3,524      | 3,492      |

**ARF & shock cohorts**

|                  |            |            |            |            |
|------------------|------------|------------|------------|------------|
| ARF, 4h          | 15,873 (18.3)| 11,147 (18.2)| 2,368 (18.1)| 2,358 (18.7)|
| ARF, 12h         | 14,174 (9.7) | 9,971 (9.7) | 2,110 (9.3) | 2,093 (9.6) |
| shock, 4h        | 19,342 (14.9)| 13,613 (14.8)| 2,862 (15.8)| 2,867 (14.7)|
| shock, 12h       | 17,588 (7.8) | 12,381 (7.6) | 2,595 (8.4) | 2,612 (7.9) |
### Table S3.3: eICU Study Cohort Characteristics.

| Population | ARF, 4h N=138,840 | ARF, 12h N=122,619 | shock, 4h N=164,333 | shock, 12h N=144,725 |
|------------|-------------------|-------------------|---------------------|---------------------|
| N=77,066   | 66 (54-77)        | 65 (53-77)        | 65 (53-77)          | 65 (52-76)          |
| Age, median (IQR) | 34,920 (45.3) | 64,950 (46.8) | 57,434 (46.8) | 76,229 (46.4) |
| Sex: F (%) |                  |                  |                     |                     |
| Race       |                   |                   |                     |                     |
| African American | 9,083 (11.8)     | 13,925 (10.0)    | 12,463 (10.2)       | 18,223 (11.1)       |
| Asian      | 1,270 (1.6)       | 2,317 (1.7)      | 2,045 (1.7)         | 2,546 (1.5)         |
| Caucasian  | 59,191 (76.8)     | 107,957 (77.8)   | 95,416 (77.8)       | 126,621 (77.1)      |
| Hispanic   | 2,853 (3.7)       | 5,527 (4.0)      | 4,919 (4.0)         | 6,127 (3.7)         |
| Native American | 449 (0.6)     | 1,189 (0.9)      | 941 (0.8)           | 1,324 (0.8)         |
| Other/Unknown | 4,220 (5.5)   | 7,925 (5.7)      | 6,835 (5.6)         | 9,492 (5.8)         |
| ICU type   |                   |                   |                     |                     |
| CCU-CTICU  | 6,571 (8.5)       | 9,982 (7.2)      | 9,249 (7.5)         | 12,196 (7.4)        |
| CSICU      | 2,531 (3.3)       | 6,293 (4.5)      | 4,739 (3.9)         | 6,231 (3.8)         |
| CTICU      | 2,826 (3.7)       | 2,504 (1.8)      | 2,275 (1.9)         | 4,462 (2.7)         |
| Cardiac ICU| 5,509 (7.1)       | 9,383 (6.8)      | 8,733 (7.1)         | 10,896 (6.6)        |
| MICU       | 7,374 (9.6)       | 10,792 (7.8)     | 9,617 (7.8)         | 14,402 (8.8)        |
| Med-Surg ICU| 40,736 (52.9)    | 82,333 (59.3)    | 71,951 (58.7)       | 93,265 (56.8)       |
| Neuro ICU  | 5,934 (7.7)       | 10,285 (7.4)     | 9,368 (7.6)         | 13,008 (7.9)        |
| SICU       | 5,585 (7.2)       | 7,268 (5.2)      | 6,687 (5.5)         | 9,873 (6.0)         |
| Outcomes   |                   |                   |                     |                     |
| Primary outcome (%) | 8859 (11.5) | 9527 (6.9) | 6942 (5.7) | 12448 (7.6) |
| LOS, median (IQR) | 88 (64-142) | 34 (20-63) | 38 (23-66) | 39 (21-71) |

**Acronyms:** CCU-CTICU, coronary care unit/cardiothoracic intensive care unit; CSICU, cardiac surgery intensive care unit; CTICU, cardiothoracic intensive care unit; Cardiac ICU, cardiac intensive care unit; MICU, medical intensive care unit; Med-Surg ICU, medical-surgical intensive care unit; Neuro ICU, neurological intensive care unit; SICU, surgical intensive care unit.

### Table S3.4: eICU Data Splitting.

For all tasks, we first assigned each patient in the source population to either of train/val/test partition with 70/15/15 ratio, and then split the study cohorts accordingly. We list the counts of patients or examples with the percentage of positive cases (in parentheses) where appropriate.

| Population          | Total   | Train   | Val     | Test    |
|---------------------|---------|---------|---------|---------|
| **Source population**|         |         |         |         |
| Patients            | 139,367 | 97,556  | 20,905  | 20,906  |
| ICU stays           | 200,859 | 140,646 | 30,137  | 30,076  |
| **In-hospital mortality, ARF & shock cohorts** | | | | |
| In-hospital mortality, 48h | 77,066 (11.5) | 53,842 (11.5) | 11,682 (12.0) | 11,542 (11.0) |
| ARF, 4h             | 138,840 (6.9) | 97,199 (6.9) | 20,892 (6.9) | 20,749 (6.6) |
| ARF, 12h            | 122,619 (5.7) | 85,845 (5.7) | 18,499 (5.7) | 18,275 (5.5) |
| shock, 4h           | 164,333 (7.6) | 115,070 (7.7) | 24,616 (7.4) | 24,647 (7.3) |
| shock, 12h          | 144,725 (4.9) | 101,393 (5.1) | 21,690 (4.6) | 21,642 (4.6) |
4. Discussion of FIDDLE Execution

For applying FIDDLE to MIMIC-III cohorts, we used a Linux machine with 500GB of RAM and 100 CPU cores. For applying FIDDLE to eICU cohorts, we used a Linux compute instance on Google Cloud Platform with 3TB of RAM and 160 CPU cores.

4.1 FIDDLE applied to MIMIC-III

When evaluating on the task of in-hospital mortality, we set the prediction time $T = 48$ hours and followed the same inclusion and exclusion criteria in Harutyunyan et al.\[5\] (although only analyzing patients in the MetaVision system due to its relative recency).

The Pre-filter step removed about a third of variables, but the number of rows in the table decreased by < 1%. Note that if we had kept these variables, their associated features would eventually be filtered out in Post-filter. By setting $\theta_1 = \theta_2$, we effectively removed these variables ahead of time, thereby avoiding wasted computation in transforming rare variables (e.g., removing a third of variables would reduce the number of variables that Transform needs to consider by a third, thereby reducing about a third of the runtime). Hence, Pre-filter is important to ensure that the pipeline runs efficiently.

The Transform step identified the following frequent variables: heart rate, respiratory rate, systolic and diastolic blood pressures, and peripheral oxygen saturation. These correspond to vital sign measurements, which we expect to be frequently and consistently recorded in the ICU. By setting $\theta_{freq} = 1$, we identified those variables that were recorded more than once per hour on average. Since some patients might not have a single recording of one of these frequent variables, not all missing values are imputed. In fact, less than half of the missing values were imputed. Notably, for a larger $T = 48$, a larger fraction of the missing values can be imputed, compared to a smaller $T = 4$. As the number of the prediction window increases, we expect more patients to have at least one recorded measurement at any time (they are frequent in expectation), thus allowing imputation forward in time.

The computational complexity of Transform is linear in $N, L$ and $D$. However, since each of the two stages contain parallelizable operations, in our implementation, we parallelized encoding each $1 \times L \times D$ slice over the $N$ examples, and variable discretization using each $N \times L \times 1$ slice over the $D$ variables.

The Post-filter step removed 30%–40% of binary features from Transform. In the resulting binary feature matrix, only about < 2% entries of the feature matrices were nonzero. This enabled efficient storage by using sparse data structures (approximately 10–200MB for >10K examples and >4K feature columns over up to 48 time slices).

4.2 FIDDLE applied to eICU

The cohorts of the eICU dataset were approximately 10 times as large as those of MIMIC-III on average. The frequent variables identified, similar to the case of MIMIC-III, pertain to important vital signs. Among the final feature matrices, the largest was for the mortality cohort (X: ~700MB) when optimized with sparse data structures.
Table S4.1: FIDDLE intermediate output and data properties for MIMIC-III.

|                      | mortality, 48h | ARF, 4h | ARF, 12h | shock, 4h | shock, 12h |
|----------------------|--------------|---------|----------|-----------|------------|
| N                    | 8,577        | 15,873  | 14,174   | 19,342    | 17,588     |
| L                    | 48           | 4       | 12       | 4         | 12         |
| Runtime              | ~2.4h        | 30 min  | 45 min   | 40 min    | 60 min     |
|                      | ~8700 sec    | ~1700 sec| ~2800 sec| ~2200 sec | ~3600 sec  |

(1) Pre-filter

|                      | Input rows   | Output rows | Total variables | Rare var | Remaining var |
|----------------------|--------------|-------------|-----------------|----------|---------------|
|                      | 33,684,409   | 33,661,000  | 5,405           | 1,524    | 3,881         |
|                      | 5,034,261    | 5,021,722   | 4,364           | 1,864    | 2,500         |
|                      | 11,967,172   | 11,948,197  | 4,730           | 1,839    | 2,891         |
|                      | 6,747,035    | 6,732,523   | 4,626           | 1,946    | 2,680         |
|                      | 16,547,927   | 16,526,561  | 5,014           | 1,899    | 3,115         |

(2) Transform

|                      | Time-invariant var | Time-invariant features | Time-dependent var | Non-freq missing | Frequent var |
|----------------------|--------------------|-------------------------|-------------------|-----------------|-------------|
|                      | 12                 | 152                     | 3,869             | 1,561,275,820   | DBP         |
|                      | 12                 | 159                     | 2,488             | 153,598,742     | HR          |
|                      | 12                 | 157                     | 2,879             | 478,738,648     | RR          |
|                      | 12                 | 164                     | 2,668             | 200,563,662     | SpO2        |
|                      | 12                 | 162                     | 3,103             | 639,797,667     | SBP         |

|                      | Non-freq missing   | Frequent var            | Freq missing      | Freq imputed    | Time-dependent features |
|----------------------|--------------------|-------------------------|------------------|----------------|-------------------------|
|                      | out of 1,590,381,648| DBP                     | 179,265          | 97,328         | 9,715                   |
|                      | out of 157,650,636 | HR                      | 84,277           | 6,397          | 6,245                   |
|                      | out of 488,832,912 | RR                      | 95,971           | 28,008         | 7,184                   |
|                      | out of 206,030,984 | SpO2                    | 100,599          | 7,402          | 6,890                   |
|                      | out of 653,851,488 | SBP                     | 119,852          | 33,960         | 7,925                   |

(3) Post-filter

|                      | time-invariant similar | time-invariant duplicated | d, final time-invariant features | time-dependent similar | time-dependent duplicated | D, final time-dependent features |
|----------------------|-------------------------|---------------------------|---------------------------------|-------------------------|---------------------------|---------------------------------|
|                      | 53                      | 3                         | 96                             | 2,068                   | 340                       | 7,307                           |
|                      | 58                      | 3                         | 98                             | 1,994                   | 206                       | 4,045                           |
|                      | 58                      | 3                         | 96                             | 2,119                   | 249                       | 4,816                           |
|                      | 63                      | 3                         | 98                             | 2,166                   | 202                       | 4,522                           |
|                      | 62                      | 3                         | 97                             | 2,181                   | 244                       | 5,500                           |
Table S4.2: FIDDLE intermediate output and data properties for eICU.

|                | mortality, 48h | ARF, 4h | ARF, 12h | shock, 4h | shock, 12h |
|----------------|----------------|---------|----------|-----------|------------|
| N              | 77,066         | 138,840 | 122,619  | 164,333   | 144,725    |
| L              | 48             | 4       | 12       | 4         | 12         |
| $\theta_1, \theta_2$ | (0.01, 0.01)  | (0.001, 0.001) | (0.01, 0.001) | (0.001, 0.001) | (0.01, 0.001) |

(1) Pre-filter

|                | Input rows     | Output rows   | Total variables | Rare var | Remaining var |
|----------------|----------------|---------------|-----------------|----------|---------------|
|                | 316,318,631    | 311,318,918   | 33,715          | 31,738   | 1,977         |
|                | 48,697,857     | 48,150,487    | 32,098          | 26,725   | 5,373         |
|                | 116,316,539    | 112,643,583   | 33,287          | 32,258   | 1,029         |
|                | 60,569,983     | 59,900,750    | 34,370          | 28,730   | 5,640         |
|                | 143,378,001    | 138,682,397   | 35,657          | 34,569   | 1,088         |

(2) Transform

|                | Time-invariant var | Time-invariant features | Time-dependent var | Time-dependent features |
|----------------|---------------------|-------------------------|--------------------|-------------------------|
|                | 37                  | 22                      | 18                 | 38                      | 29                      |
|                | 79,881              | 127,693                 | 118,367            | 148,580                 | 136,898                 |
|                | 5,351               | 1,011                   | 5,602              | 1,059                   |
|                | 27,193              | 39,703                  | 10,744             | 43,958                  | 12,372                  |

(3) Post-filter (% density)

|                | $d$, final time-invariant features | $D$, final time-dependent features |
|----------------|-----------------------------------|-----------------------------------|
|                | 146 (7.9)                         | 48 × 2,382 (2.8)                  |
|                | 717 (1.7)                         | 4 × 5,854 (1.3)                   |
|                | 119 (9.0)                         | 12 × 2,713 (2.6)                  |
|                | 770 (1.6)                         | 4 × 6,314 (1.2)                   |
|                | 128 (8.5)                         | 12 × 2,946 (2.4)                  |
5. Hyperparameters and Model Selection

For MIMIC-III study cohort corresponding to the task of predicting in-hospital mortality, we used the same train-test splits as in Harutyunyan et al. [5], except that we excluded the subset of patients not recorded in the MetaVision system. For all other cohorts, a randomly selected 15% of examples were held out, and the remaining 85% were used for training and model selection.

For the LR and RF models, we selected hyperparameters (Table S5.1, first two rows) via 5-fold cross validation on the training data using a random search with a budget of 50, maximizing the average area under the receiver operating characteristics curve (AUROC) across folds. After hyperparameter selection, we retrained the model on the entire training set. On MIMIC-III, we tuned the hyperparameters for the CNN and LSTM models (Table S5.1, last two rows) using a random search with the same budget of 50. On eICU, we used a budget of 5 due to the significantly larger sample size. We trained the models for a maximum of 15 epochs, selecting the model and hyperparameter setting that led to the best performance (AUROC) on a held-out portion of the training data (i.e., the validation set).

Table S5.1: Hyperparameter search space for each model type.

| Model | Description | Training Hyperparameters | Model Hyperparameters |
|-------|-------------|--------------------------|-----------------------|
| LR    | L2-regularized Logistic Regression | - | \( C \): inverse of regularization strength | \( C \in \text{LogUniform}[-5, 5] \) |
| RF    | Random Forest | - | criterion, max_depth, max_features, min_samples_leaf, min_samples_split, n_estimators | \{gini, entropy\}, \{4, 8, 16, 32, None\}, Uniform\{1, 100\}, Uniform\{2, 11\}, Uniform\{1, 11\}, Uniform\{50, 500\} |
| CNN   | Series of 1D convolutions followed by dense layers | \( \eta \): learning rate, \( B \): batch size | \( W \): width of conv filters, \( K \): number of conv layers, \( d_c \): number of conv filters in each layer, \( d_f \): number of fully-connected layer neurons, \( p \): dropout rate, \( f \): activation function | \( W \in \{1, 2, 3, 4\} \), \( K \in \{1, 2\} \), \( d_c \in \{16, 32, 64, 128\} \), \( d_f \in \{16, 32, 64, 128\} \), \( p \in \{0, 0.1, 0.2, 0.4, 0.8\} \), \( f \in \{\text{relu, elu}\} \) |
| LSTM  | Single- or multi-layer LSTM network, followed by dense layers | \( \eta \in \{1e-2, 1e-3, 1e-4\} \), \( B \in \{16, 32, 64, 128\} \) | \( H \): hidden size, \( K \): number of LSTM layers, \( d_f \): number of fully-connected layer neurons, \( p \): dropout rate, \( f \): activation function | \( H \in \{16, 32, 64, 128\} \), \( K \in \{1, 2, 3\} \), \( d_f \in \{16, 32, 64, 128\} \), \( p \in \{0, 0.1, 0.2, 0.4, 0.8\} \), \( f \in \{\text{relu, elu}\} \) |
6. Details of MIMIC-Extract

We used the implementation provided by the authors of MIMIC-Extract [13] at https://github.com/MLforHealth/MIMIC_Extract (as of September 2019). To facilitate comparison, we made a few changes to the pipeline. First, we adjusted MIMIC-Extract to ensure that we were extracting every ICU record (and treating each one as an example), because MIMIC-Extract only looks at the first ICU visit of each patient. Second, we set all MIMIC-Extract arguments to the default values except for the following:

```
--min_percent 0 \
--min_duration 0 \
--max_duration 999999
```

These arguments ensure that the extracted population is a superset of each study cohort population defined in Table S3.1.

Once the data-frames were created using MIMIC-Extract, we followed the provided notebook for in-hospital mortality prediction to create feature matrices for each of our cohorts, and then applied the same training and evaluating procedure with the same inclusion/exclusion criteria and the same train/test splits as reported in the main text. For all of our tasks, we set gap time = 0. In order to construct the time dependent tensor $\mathcal{X}$, we looked at the vitals_hourly_data and outcomes_hourly_data dataframes. (outcomes_hourly_data stores data about interventions and is not originally used by the provided notebook to make mortality predictions). For the vitals_hourly_data, as suggested in the notebook, we performed simple imputation for each train, test, and validation set separately using our predefined partition sets. In addition to normalizing the mean value for each variable, we also normalized the time_since_last_measured feature for each column (which was not done in the original notebook). It must be noted that before normalizing, we set all the standard deviations that were 0 to np.nan, because otherwise the dataframes would contain inf values without raising any errors. We then concatenated labs_vitals with outcomes_hourly_data and reshaped it into an $N \times L \times D$ tensor. We also used the time independent (i.e. static) variables extracted by MIMIC_Extract. Originally, the mortality prediction notebook only uses the time-invariant variables to construct the labels for each example, but we also used them for prediction. We used 'age', 'ethnicity', 'gender', and 'first_careunit' (which are the features they used in their notebook to make intervention predictions). These time-invariant variables gave us $\mathbf{S}$ of dimension $N \times d̂$. 
7. Extended Experimental Results

7.1 MIMIC-III with discretize = ‘false’

After extracting the features, we normalized the output features to be between 0 and 1 and evaluated on the same tasks and ML models. Compared to results reported in Table S7.1, ML models trained using discretized features achieved higher AUROC and AUPR scores in general (Table 5A in main text).

Table S7.1: Model performance of ARF and shock tasks, in terms of AUROC & AUPR on the held-out test set with 95% confidence intervals.

| Task                      | in-hospital mortality (48h) N=11,542 | ARF (4h) N=20,749 | ARF (12h) N=18,275 | shock (4h) N=24,647 | shock (12h) N=21,642 |
|---------------------------|--------------------------------------|--------------------|---------------------|----------------------|-----------------------|
| Method                    | AUROC 0.853 (0.818, 0.886) | AUPR 0.445 (0.352, 0.538) | AUROC 0.798 (0.774, 0.820) | AUPR 0.622 (0.582, 0.662) | AUROC 0.735 (0.701, 0.770) | AUPR 0.264 (0.210, 0.323) | AUROC 0.797 (0.773, 0.821) | AUPR 0.505 (0.454, 0.555) | AUROC 0.757 (0.724, 0.791) | AUPR 0.238 (0.188, 0.296) |
| FIDDLE-LR                 | 0.823 (0.786, 0.856) | 0.376 (0.298, 0.470) | 0.815 (0.793, 0.838) | 0.653 (0.613, 0.692) | 0.764 (0.728, 0.794) | 0.320 (0.266, 0.391) | 0.823 (0.800, 0.845) | 0.547 (0.497, 0.593) | 0.804 (0.772, 0.836) | 0.339 (0.276, 0.409) |
| FIDDLE-CNN                | 0.868 (0.832, 0.903) | 0.513 (0.423, 0.616) | 0.794 (0.770, 0.817) | 0.622 (0.582, 0.663) | 0.734 (0.698, 0.770) | 0.271 (0.214, 0.340) | 0.795 (0.771, 0.818) | 0.489 (0.437, 0.538) | 0.745 (0.710, 0.778) | 0.231 (0.184, 0.291) |
| FIDDLE-LSTM               | 0.867 (0.828, 0.901) | 0.509 (0.418, 0.601) | 0.805 (0.781, 0.826) | 0.629 (0.588, 0.667) | 0.757 (0.725, 0.790) | 0.287 (0.230, 0.349) | 0.789 (0.764, 0.813) | 0.485 (0.434, 0.538) | 0.753 (0.717, 0.785) | 0.250 (0.206, 0.307) |

7.2 Performance of all FIDDLE-based ML models

On both datasets, all four types of ML models achieved good discriminative performance and exhibited good model calibration on the five prediction tasks. We display the ROC curves, PR curves, and calibration curves in Table S7.2 (for MIMIC-III) and Table S7.3 (for eICU).
Table S7.2: Evaluation of FIDDLE-based models on MIMIC-III.

| Task          | ROC curve | PR curve | Calibration curve |
|---------------|-----------|----------|-------------------|
| **Mortality 48h** | ![ROC curve](image1) | ![PR curve](image2) | ![Calibration curve](image3) |
| **ARF 4h**    | ![ROC curve](image4) | ![PR curve](image5) | ![Calibration curve](image6) |
| **ARF 12h**   | ![ROC curve](image7) | ![PR curve](image8) | ![Calibration curve](image9) |
| **Shock 4h**  | ![ROC curve](image10) | ![PR curve](image11) | ![Calibration curve](image12) |
| **Shock 12h** | ![ROC curve](image13) | ![PR curve](image14) | ![Calibration curve](image15) |
Table S7.3: Evaluation of FIDDLE-based models on eICU.

| Task          | ROC curve | PR curve | Calibration curve |
|---------------|-----------|----------|-------------------|
| Mortality 48h | ![ROC curve](image1.png) | ![PR curve](image2.png) | ![Calibration curve](image3.png) |
| ARF 4h        | ![ROC curve](image4.png) | ![PR curve](image5.png) | ![Calibration curve](image6.png) |
| ARF 12h       | ![ROC curve](image7.png) | ![PR curve](image8.png) | ![Calibration curve](image9.png) |
| Shock 4h      | ![ROC curve](image10.png) | ![PR curve](image11.png) | ![Calibration curve](image12.png) |
| Shock 12h     | ![ROC curve](image13.png) | ![PR curve](image14.png) | ![Calibration curve](image15.png) |
7.3 Comparing FIDDLE-based models to additional baselines for MIMIC-III

In-hospital mortality. We compared to the benchmark by Harutyunyan et al. [5], using the implementation at [https://github.com/YerevaNN/mimic3-benchmarks](https://github.com/YerevaNN/mimic3-benchmarks). As reported by Harutyunyan et al., we used the following command and hyperparameter setting to train the model, and stopped training after 28 epochs when best validation performance was achieved:

```
python -um mimic3models.in_hospital_mortality.main \
   --network mimic3models/keras_models/lstm.py \
   --dim 16 --timestep 1.0 --depth 2 --dropout 0.3 --mode train --batch_size 8 \
   --output_dir mimic3models/in_hospital_mortality
```

ARF and shock. We compared to the National Early Warning Score (NEWS) [15], a rule-based severity scoring system that considers instantaneous values of important vital signs, consciousness and supplemental oxygen. For each ICU stay, we calculated NEWS every time a new measurement was made (values of other unmeasured NEWS components were handled by carry-forward imputation). To summarize these scores, we took the maximum leading up to the prediction time, because this resembles how physicians will likely react in practice [16].

Among the four models using FIDDLE features, three (LR, CNN, and LSTM) performed significantly better than the benchmark LSTM model when evaluated on the same held-out test set ($p$-value < 0.001, Table S7.4). For prediction of both ARF and shock at 4 hours after ICU admission, all four ML models performed significantly better than NEWS in terms of AUROC ($p$-values < 0.001; Table S7.5). Note that all comparisons with NEWS were only performed on subsets of the held-out test set for which NEWS could be calculated.

### Table S7.4: Model performance on the mortality prediction task compared to the benchmark LSTM [5].

| Method          | AUROC  | AUPR   |
|-----------------|--------|--------|
| Benchmark LSTM  | 0.839  | 0.492  |
| FIDDLE-LR       | 0.856  | 0.444  |
| FIDDLE-RF       | 0.814  | 0.357  |
| FIDDLE-CNN      | 0.886  | 0.531  |
| FIDDLE-LSTM     | 0.868  | 0.510  |

### Table S7.5: Model performance on the ARF and shock prediction tasks compared to NEWS.

| Method          | ARF (4h) | ARF (12h) | shock (4h) | shock (12h) |
|-----------------|----------|-----------|------------|-------------|
|                 | N=1,823  | N=1,950   | N=2,233    | N=2,429     |
| NEWS            | 0.650    | 0.628     | 0.677      | 0.682       |
|                 | (0.614, 0.687) | (0.588, 0.666) | (0.644, 0.871) | (0.643, 0.721) |
| FIDDLE-LR       | 0.733    | 0.755     | 0.775      | 0.793       |
|                 | (0.699, 0.767) | (0.717, 0.789) | (0.745, 0.805) | (0.758, 0.826) |
| FIDDLE-RF       | 0.739    | 0.759     | 0.755      | 0.773       |
|                 | (0.703, 0.772) | (0.722, 0.793) | (0.725, 0.789) | (0.738, 0.807) |
| FIDDLE-CNN      | 0.750    | 0.768     | 0.788      | 0.795       |
|                 | (0.718, 0.783) | (0.732, 0.801) | (0.761, 0.817) | (0.763, 0.826) |
| FIDDLE-LSTM     | 0.744    | 0.767     | 0.777      | 0.794       |
|                 | (0.710, 0.777) | (0.732, 0.800) | (0.747, 0.808) | (0.761, 0.826) |
7.4 Inspection of learned models

EHR features derived from the FIDDLE preprocessing pipeline can be interpreted by clinicians and those familiar with how the data were collected. When used in a linear model, one can easily identify which features are associated with either an increase or decrease in estimated patient’s risk (Table S7.6). Alternatively, one can use feature importance metrics like permutation importance to identify what features are driving prediction. In our analysis, some predictive factors were specific to each task, for example, lower blood pH is correlated with higher risk of ARF, while lower systolic blood pressure is correlated with higher risk of shock. Other predictive factors were shared across tasks, for example, being located in cardiac surgery recovery unit increased risk, while being located in a medical intensive care unit conferred lower risk. Such analysis enables us to check if the models produce sensible predictions and allows us to detect potential label leakage. We encourage users of FIDDLE to carefully review the features produced by FIDDLE and, if possible, the feature importances. Though FIDDLE is a data-driven approach and can help speed up analysis, it cannot replace the critical step of checking and debugging the model.

Table S7.6: Ranked feature coefficients from logistic regression models, for the two 4h prediction tasks.

| Feature                                           | Coef. | Feature                                           | Coef. |
|---------------------------------------------------|-------|---------------------------------------------------|-------|
| Cardiac Surgery Recovery Unit                     | 0.38  | Cardiac Surgery Recovery Unit                     | 0.28  |
| ICU Ward ID = 15                                  | 0.20  | ICU Ward ID = 15                                  | 0.17  |
| ICU Ward ID = 12                                  | 0.18  | Age in Q3 (60, 69]                                | 0.12  |
| at 2-3h, pH in Q1 (6.55, 7.3]                      | 0.17  | Transfer from Hospital                            | 0.12  |
| at 3-4h, pH in Q1 (6.55, 7.3]                      | 0.16  | ICU Ward ID = 12                                  | 0.11  |
| at 2-3h, Base Excess in Q1 (< –4)                 | 0.16  | at 2-3h, pH in Q1 (6.55, 7.3]                      | 0.10  |
| at 3-4h, Foley catheter output in Q1 (< 40)       | 0.14  | at 3-4h, Foley catheter output in Q1 (< 40)       | 0.10  |
| at 2-3h, Blood Glucose in Q5 (22.9, 110]           | 0.14  | Age in Q4 (69, 81]                                | 0.10  |
| Transfer from Hospital                            | 0.14  | at 3-4h, Invasive Ventilation                     | 0.10  |
| Age in Q3 (61, 67]                                 | 0.14  | at 3-4h, min systolic BP in Q1 (< 101)            | 0.09  |
| at 3-4h, Heart Rhythm = Sinus Rhythm              | -0.12 | at 1-2h, cough effort = strong                    | -0.08 |
| at 3-4h, Heart rate recorded                      | -0.12 | Ethnicity = Black                                 | -0.09 |
| at 1-2h, O2 Delivery Device = None                | -0.12 | at 3-4h, Heart Rhythm = Sinus Rhythm              | -0.09 |
| at 2-3h, Head of Bed = 30 degrees                 | -0.13 | Admission to ICU time in Q4 (1.2h, 32h]           | -0.09 |
| Age in Q5 (> 81)                                  | -0.15 | at 3-4h, Heart Rate recorded                      | -0.10 |
| at 2-3h, O2 Delivery Device = None                | -0.15 | at 1-2h, Admit from ED                            | -0.10 |
| ICU Ward ID = 52                                  | -0.16 | at 3-4h, min systolic BP in Q5 (> 141)            | -0.13 |
| at 1-2h, Admit from ED                            | -0.18 | ICU Ward ID = 52                                  | -0.13 |
| Coronary Care Unit; ICU Ward ID = 7               | -0.18 | Age in Q1 (18, 49]                                | -0.17 |
| Medical ICU                                       | -0.24 | Medical ICU                                       | -0.21 |
7.5 Sensitivity Analyses & Flexibility of FIDDLE

To understand the effect of user-defined arguments on the utility of the features generated by FIDDLE, we tested FIDDLE using (i) different filtering thresholds, $\theta = \theta_1 = \theta_2$ and (ii) temporal granularities, $dt$. In addition, to explore alternatives to FIDDLE’s default solution for representing numerical variables and handling missing data (Table 1 in main text), we made two additional sets of comparisons: (iii) a continuous vs. one-hot encoding vs. ordinal encoding representation, and (iv) carry-forward imputation vs. median imputation vs. no imputation. These four sets of experiments also serve to demonstrate that FIDDLE is flexible and can be adapted based on the specific needs of the end user or the task. We performed the above comparisons in the context of predicting in-hospital mortality using MIMIC-III and report the results below.

7.5.1 Sensitivity Analyses of FIDDLE Arguments

When FIDDLE was run with different values of $\theta$ (Table S7.7), more aggressive filtering (larger $\theta$) led to faster runtime and a smaller number of features, with only slight drop in performance. Overall, performance of ML models was relatively robust across different amounts of filtering. This is in part because the filtering applied to time-dependent features is quite conservative; FIDDLE only removes features that are constant over all time steps for most patients, retaining important temporal trends.

When we varied the value of $dt$ (Table S7.8), smaller $dt$ values resulted in more time bins (larger $L$) and gave deep models (CNN and LSTM) a slight advantage over non-deep approaches (LR and RF). We hypothesize this advantage is due, in part, to the ability of the CNN and LSTM to capture temporal patterns across time-steps. When $dt$ is large, FIDDLE recognizes more variables as “frequent” and summarizes their values according to the calculated summary statistics, resulting in more time-dependent features ($D$) for each time-step. These features improve the performance of non-deep models, reducing the performance gap between deep and non-deep approaches. In practice, provided sufficient computational resources, one could tune $dt$ as a hyperparameter and optimize it for validation performance depending on the choice of modeling techniques.

7.5.2 Flexibility of FIDDLE

Among the different representations of numerical variables (Table S7.9), the two discretization approaches (i.e., one-hot and ordinal encodings) slightly outperformed the approach without discretization for all models except for RF (where they performed worse), though the differences are not significant. Similarly, the method of imputation did not have a significant impact on the performance of downstream ML models (Table S7.10). Given the open-source nature of the software, future researchers may consider alternatives to FIDDLE’s defaults and select the most appropriate representation depending on their specific applications.

### Table S7.7: AUROC scores on the held-out test set (N=1,264) for in-hospital mortality prediction at 48h, with different filtering thresholds, $\theta$ (for $dt = 1$).

| $\theta$ | 0.001  | 0.01  | 0.05  | 0.1   | 0.2   | 0.4   |
|----------|--------|-------|-------|-------|-------|-------|
| Processing time | ~2.4h  | ~2h   | ~1h   | ~1h   | ~30min | ~20min |
| $d$  | 97     | 57    | 43    | 34    | 16    | 4     |
| $D$  | 7,411  | 3,872 | 1,969 | 1,398 | 902   | 522   |
| LR    | 0.856  | 0.855 | 0.854 | 0.853 | 0.848 | 0.830 |
|       | (0.821, 0.888) | (0.820, 0.887) | (0.818, 0.886) | (0.818, 0.886) | (0.811, 0.881) | (0.793, 0.865) |
| RF    | 0.814  | 0.814 | 0.814 | 0.814 | 0.814 | 0.794 |
|       | (0.780, 0.847) | (0.780, 0.847) | (0.778, 0.848) | (0.777, 0.849) | (0.779, 0.847) | (0.755, 0.831) |
| CNN   | 0.886  | 0.875 | 0.874 | 0.880 | 0.878 | 0.866 |
|       | (0.854, 0.916) | (0.842, 0.905) | (0.840, 0.908) | (0.850, 0.908) | (0.843, 0.910) | (0.832, 0.898) |
| LSTM  | 0.868  | 0.871 | 0.866 | 0.866 | 0.876 | 0.869 |
|       | (0.835, 0.897) | (0.838, 0.901) | (0.836, 0.894) | (0.831, 0.897) | (0.843, 0.906) | (0.836, 0.899) |
Table S7.8: AUROC scores on the held-out test set (N=1,264) for in-hospital mortality prediction at 48h, with different temporal granularities, $dt$ (for $\theta = 0.001$).

| $dt$ | 1  | 4  | 12 | 24 | 48 |
|------|----|----|----|----|----|
| **Processing time** | ~2.4h | ~1.5h | ~50min | ~40min | ~20min |
| $d$ | 96 | 96 | 96 | 96 | 96 |
| $D$ | 7,307 | 7,418 | 7,925 | 8,331 | 8,328 |
| $L$ | 48 | 12 | 4 | 2 | 1 |
| $d + LD$ | 350,832 | 89,112 | 31,796 | 16,758 | 8,424 |

| Model | Continuous | One-hot encoding | Ordinal encoding |
|------|-------------|-----------------|-----------------|
| **LR** | 0.856 (0.821, 0.888) | 0.887 (0.853, 0.915) | 0.897 (0.866, 0.924) | 0.908 (0.880, 0.932) | 0.900 (0.869, 0.927) |
| **RF** | 0.814 (0.780, 0.847) | 0.850 (0.818, 0.879) | 0.876 (0.848, 0.901) | 0.888 (0.860, 0.911) | 0.889 (0.861, 0.913) |
| **CNN** | 0.886 (0.854, 0.916) | 0.878 (0.843, 0.907) | 0.894 (0.861, 0.923) | 0.899 (0.868, 0.926) | (see note) |
| **LSTM** | 0.868 (0.835, 0.897) | 0.887 (0.853, 0.916) | 0.898 (0.869, 0.924) | 0.899 (0.867, 0.926) | 0.889 (0.858, 0.918) |

Note: The result for $dt = 48$ using a CNN is omitted because the pooling operation in a CNN does not apply to a single time-step.

Table S7.9: AUROC scores on the held-out test set (N=1,264) for in-hospital mortality prediction at 48h, with different representations of numerical variables (for $dt = 1, \theta = 0.001$).

| Representation | Continuous | One-hot encoding | Ordinal encoding |
|----------------|-------------|-----------------|-----------------|
| $d$ | 88 | 96 | 94 |
| $D$ | 3,941 | 7,307 | 7,949 |
| **LR** | 0.853 (0.818, 0.886) | 0.856 (0.821, 0.888) | 0.861 (0.829, 0.892) |
| **RF** | 0.823 (0.786, 0.856) | 0.814 (0.780, 0.847) | 0.814 (0.777, 0.848) |
| **CNN** | 0.868 (0.832, 0.903) | 0.886 (0.854, 0.916) | 0.885 (0.854, 0.912) |
| **LSTM** | 0.867 (0.828, 0.901) | 0.868 (0.835, 0.897) | 0.885 (0.856, 0.911) |
Table S7.10: AUROC scores on the held-out test set (N=1,264) for in-hospital mortality prediction at 48h, with different imputation methods (for a range of \( dt \) values and \( \theta = 0.1 \)).

| Representation | \( dt = 1 \) | \( dt = 4 \) | \( dt = 12 \) |
|----------------|--------------|--------------|-------------|
|                | Carry-forward impute | Median impute | No impute  | Carry-forward impute | Median impute | No impute  | Carry-forward impute | Median impute | No impute  |
| \( d \)        | 34           | 34           | 34         | 34           | 34           | 34         | 34           | 34           | 34         |
| \( D \)        | 1,398        | 1,397        | 1,397      | 1,545        | 1,536        | 1,528      | 1,938        | 1,936        | 1,841      |
| LR             | 0.853 (0.818, 0.886) | 0.852 (0.817, 0.885) | 0.854 (0.819, 0.886) | 0.884 (0.851, 0.914) | 0.885 (0.851, 0.914) | 0.882 (0.848, 0.911) | 0.893 (0.861, 0.921) | 0.895 (0.864, 0.923) | 0.891 (0.861, 0.918) |
| RF             | 0.813 (0.776, 0.846) | 0.821 (0.786, 0.853) | 0.814 (0.779, 0.848) | 0.863 (0.830, 0.892) | 0.861 (0.828, 0.889) | 0.861 (0.829, 0.889) | 0.883 (0.853, 0.909) | 0.880 (0.848, 0.908) | 0.883 (0.854, 0.909) |
| CNN            | 0.884 (0.849, 0.914) | 0.881 (0.847, 0.913) | 0.881 (0.846, 0.913) | 0.857 (0.824, 0.890) | 0.891 (0.858, 0.920) | 0.884 (0.852, 0.912) | 0.884 (0.853, 0.913) | 0.895 (0.863, 0.922) | 0.886 (0.856, 0.913) |
| LSTM           | 0.866 (0.831, 0.897) | 0.868 (0.834, 0.898) | 0.867 (0.833, 0.898) | 0.884 (0.852, 0.913) | 0.885 (0.852, 0.914) | 0.890 (0.858, 0.916) | 0.902 (0.872, 0.927) | 0.892 (0.864, 0.918) | 0.895 (0.865, 0.921) |
8. Applying FIDDLE to Additional Prediction Tasks

8.1 Seq-to-seq LSTM model

Instead of a single output at the final time-step of the prediction window (as are the tasks considered in the main text), one could formulate the same task with multiple outputs, one at every time-step. Here, we demonstrate how one could adapt FIDDLE features for such modification on the task of in-hospital mortality prediction at 48h.

Since the input features remain the same, the only change is to the ground-truth labels. We use the target replication approach: if a patient’s label at the final time-step is 1 (or 0), then the label at every time-step is 1 (or 0, respectively). During evaluation, we took the maximum predicted scores over all time-steps, because this approach more accurately reflects how clinicians react to an early warning system in practice – they will act as soon as the predicted risk exceed a threshold. This inherently represents a more difficult task than the single-prediction formulation. First, the predictions at non-terminal time-steps can only make use of information up to that time-point, because we do not want any information leakage from the future. Second, if the model predicts a single false positive before true negatives, the clinicians would have already acted upon the false positive prediction.

Using the evaluation method described above, the seq-to-seq LSTM model achieved an AUROC of 0.849 (95%CI: 0.815, 0.880) and AUPR of 0.413 (95%CI: 0.328, 0.505) on the test set (N=1,264).

![ROC curve](image1.png)  ![Precision-recall curve](image2.png)

**Figure S8.1: Performance of the seq-to-seq LSTM model.**

8.2 Predicting 90-day post-discharge mortality

To test the ability of FIDDLE to handle hierarchical variables, we considered a task of predicting 90-day survival post-discharge. We treated each hospital admission in MIMIC-III (rather than each ICU stay) as an example to the ML model. We identified whether each patient has survived the hospitalization through the ‘HOSPITAL_EXPIRE_FLAG’ column in the ADMISSIONS table. We then compared patient’s recorded date of death (‘DOD’ column in PATIENTS table), with the date of discharge (‘DISCHTIME’ column in ADMISSIONS table) to determine whether the patient died within 90 days after discharge, assigning a label of 1, and a label of 0 otherwise for patients who have no recorded date of death or died after 90 days after discharge. This resulted in the cohort in **Figure S8.2.**
We extracted each patient’s discharge ICD codes from the `DIAGNOSIS_ICD` table, as well as clinical data from the following tables: `CHARTEVENTS`, `DATETIMEEVENTS`, `LABEVENTS`, `MICROBIOLOGYEVENTS`, `OUTPUTEVENTS`. We also extracted static/demographic information from `PATIENTS`, `ADMISSIONS`, and `ICUSTAYS`. Note that some of these clinical data are collected throughout the hospital admission (`LABEVENTS`) since they are sourced from the hospital database, while others are only collected within ICUs through the MetaVision system (e.g., `CHARTEVENTS`). When applying FIDDLE, we used the following settings: $T = 4,320$, $dt = 4,320$, $\theta_1 = \theta_2 = 0.01$, keeping other values as default. The time period $T$ is chosen to be 180 days (i.e., 4,320 hours) such that it is longer than the longest hospital stay, allowing FIDDLE to consider all information within each hospital admission. We set $dt = T$ so that there is only a single time bin for time-dependent data so as to speed up the processing.

To encode ICD9 codes, we considered three levels of the hierarchy/taxonomy. For example, the code “008.45” (Intestinal infection due to Clostridium difficile) corresponds to the following levels:
1. 001-139 (1. Infectious and parasitic diseases)
2. 001-009 (intestinal infectious diseases)
3. 008 (Intestinal infections due to other organisms)

We compared using different level(s) of the hierarchy, using clinical data alone, and using a combination of clinical data and diagnosis codes. In each case, the time-invariant and time-dependent features are concatenated to create a single feature vector for each example and input to LR and RF models. We summarize the predictive performance in Table S8.3.

In general, we observed that using more levels of the hierarchy led to better performance. The addition of clinical data also led to better performance for both models, outperforming that used only clinical data, suggesting that clinical data and diagnosis codes provide complementary information that can be used for the prediction.

### Table S8.3: AUROC scores on the test set (N=2,928) for 90-day post-discharge mortality prediction, using different input feature sets.

| Input features        | Dimensions | Test AUROC scores, N=2,928 |
|-----------------------|------------|-----------------------------|
|                       |            | LR                          | RF                              |
| ICD codes {1}         | 17         | 0.725 (0.694, 0.753)         | 0.733 (0.703, 0.763)            |
| ICD codes {1,2}       | 112        | 0.777 (0.748, 0.804)         | 0.778 (0.751, 0.806)            |
| ICD codes {1,2,3}     | 314        | 0.795 (0.767, 0.820)         | 0.799 (0.771, 0.825)            |
| Clinical              | 4,489      | 0.834 (0.810, 0.858)         | 0.816 (0.791, 0.842)            |
| Clinical + ICD codes {1} | 4,506    | 0.844 (0.822, 0.867)         | 0.831 (0.805, 0.854)            |
| Clinical + ICD codes {1,2} | 4,601    | 0.850 (0.827, 0.872)         | 0.846 (0.821, 0.870)            |
| Clinical + ICD codes {1,2,3} | 4,803 | 0.851 (0.829, 0.873)         | 0.852 (0.828, 0.874)            |
## 9. Qualitative Comparisons of FIDDLE to Other EHR Data Pipelines

| Pipeline             | Description                                                                                                                                                                                                 |
|----------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| “Yale Pipeline” [17] | **Overview:**  
  - Yale New Haven Health Baikal Data Science Platform, an integrated data lake and analytics platform, supports continuous patient monitoring and real-time laboratory analytics at Yale New Haven Hospital (YNHH)  
  - hardware (Hadoop) and software (Docker) solutions to support big-data workloads  
  **Input:**  
  - data elements from bedside monitors, ventilators, laboratory orders and results  
  **Output:**  
  - serialized and compressed data suitable for long-term storage in traditional data warehouse  
  **Experimental cohort:**  
  - N/A (no empirical evaluation of pipeline’s utility with respect to ML tasks was performed)  
  **Applicable use cases:**  
  - future patients at YNHH  
  **Limitations:**  
  - single site (YNHH)  
  - does not natively support the needs of downstream ML analysis |
| Pythia [18]         | **Overview:**  
  - an automated, clinically curated surgical data pipeline and repository housing all surgical patient EHR data from Duke University Health System (DUHS)  
  - designed to both promote the development of ML models and bridge the translational gap to enable rapid deployment of validated models  
  **Input:**  
  - patient clinical and surgical data across 37 million clinical encounters from the EHR  
  **Output:**  
  - every patient in the cohort represented by 194 curated clinical features, including: patient demographics (e.g., age, sex, race), smoking status, medications, comorbidities, procedure information, and proxies for surgical complexity  
  **Experimental cohort:**  
  - surgical patients from DUHS  
  **Applicable use cases:**  
  - retrospective ML analysis of any cohort within surgical patients at DUHS  
  **Limitations:**  
  - single site (DUHS)  
  - does not handle missing data (deleted all examples with missing data)  
  - currently does not adapt to rapidly changing clinical practice trends and standards of data collection within the EHR  
  - uses as snapshot representation, ignoring temporal trends |
| MIMIC-Extract [13]  | **Overview:**  
  - an open-source pipeline for transforming raw EHR data for critical care patients contained in the publicly available MIMIC-III database into data-frames that are usable for common machine learning models  
  **Input:**  
  - MIMIC-III database v1.4  
  **Output:**  
  - 100+ clinically aggregated variables for each patient, including: 10 time-invariant demographic variables, and 93 time-varying variables related to vitals and labs, binned hourly |
| **Experimental cohort:** | **“FHIR Pipeline” by Rajkomar et al. [12]** |
|-------------------------|------------------------------------------|
| • validated on 34,472 ICU stays in MIMIC-III, which are the first ICU admissions for adult patients (i.e. age > 15), whose length of stay ≥12 hours and <10 days | **Overview:** |
| | • a representation of patients’ entire raw EHR records based on the FHIR format generated through deep learning methods |
| **Applicable use cases:** | **Input:** |
| • retrospective ML analysis of any cohort in MIMIC-III | • EHR data in FHIR format |
| **Limitations:** | **Output:** |
| • dataset specific (MIMIC-III), thus applying to a different dataset requires redoing a lot of manual variable curation | • float-valued embedding vector for each token (e.g., medication codes, or words in a note),
| | o randomly initialized and adjusted during the training process
| | o embedding vectors for all tokens are concatenated to create the final feature vector at each timestep |
| **Experimental cohort:** | **Experimental cohort:** |
| • validated separately on two academic hospitals, each with a general patient population (not restricted to ICU) | • MIMIC-III and eICU for tasks involving 3 ICU-related outcomes |
| **Applicable use cases:** | **Applicable use cases:** |
| • retrospective ML analysis of any EHR dataset | • retrospective ML analysis of any EHR dataset |
| **Limitations:** | **Limitations:** |
| • not open-source and relies on Google’s internal systems | • not open-source and relies on Google’s internal systems
| | • deep learning techniques often require large amounts of data to learn good representations |

| **FIDDLE (this work)** |
|-------------------------|
| **Overview:** |
| • a flexible, data-driven preprocessing pipeline to transform data extracted from the EHR into feature vectors used for ML |
| **Input:** |
| • data extracted from the EHR and converted into the 4-column format |
| **Output:** |
| • feature vectors containing time-invariant and time-dependent features |
| **Experimental cohort:** |
| • MIMIC-III and eICU for tasks involving 3 ICU-related outcomes |
| **Applicable use cases:** |
| • retrospective ML analysis of any EHR dataset |
| **Limitations:** |
| • currently not integrated with the EHR; users must extract data from the EHR and format the input appropriately |
SUPPLEMENTAL REFERENCES

1. Che Z, Purushotham S, Cho K, Sontag D, Liu Y. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports* 2018;8(1):6085. doi:10.1038/s41598-018-24271-9

2. Rubin DB. Inference and missing data. *Biometrika* 1976;63(3):581-92. doi:10.1093/biomet/63.3.581

3. Nemati S, Ghassemi MM, Clifford GD. Optimal medication dosing from suboptimal clinical examples: A deep reinforcement learning approach. *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* 2016:2978-81. doi:10.1109/EMBC.2016.7591355

4. Little R, Rubin D. *Statistical analysis with missing data*. J. Wiley, 1987.

5. Harutyunyan H, Khachatrian H, Kale DC, Ver Steeg G, Galstyan A. Multitask learning and benchmarking with clinical time series data. *Scientific Data* 2019;6(1):96. doi:10.1038/s41597-019-0103-9

6. Stefan MS, Shieh M-S, Pekow PS, et al. Epidemiology and outcomes of acute respiratory failure in the United States, 2001 to 2009: a national survey. *Journal of Hospital Medicine* 2013;8(2):76-82. doi:10.1002/jhm.2004

7. Meduri GU, Turner RE, Abou-Shala N, Wunderink R, Tolley E. Noninvasive positive pressure ventilation via face mask: First-line intervention in patients with acute hypercapnic and hypoxemic respiratory failure. *Chest* 1996;109(1):179-93. doi:10.1378/chest.109.1.179

8. Gaieski DF, Mikkelsen ME. Definition, classification, etiology, and pathophysiology of shock in adults. Retrieved July, 2019, from https://www.uptodate.com/contents/definition-classification-etiology-and-pathophysiology-of-shock-in-adults.

9. Avni T, Lador A, Lev S, Leibovici L, Paul M, Grossman A. Vasopressors for the treatment of septic shock: Systematic review and meta-analysis. *PLOS One* 2015;10(8):e0129305. doi:10.1371/journal.pone.0129305

10. O'Malley KJ, Cook KP, Price MD, Wildes KR, Hurdle JF, Ashton CM. Measuring Diagnoses: ICD Code Accuracy. *Health Services Research* 2018;5:180178. doi:10.1038/sdata.2018.178

11. The Royal College of Physicians. *National Early Warning Score (NEWS): standardising the assessment of acute-illness severity in the NHS*. London: Report of a working party, 2012. https://www.rcplondon.ac.uk/projects/outputs/national-early-warning-score-news-2

12. Smith GB, Prytherch DR, Meredith P, Schmidt PE, Featherstone PI. The ability of the National Early Warning Score (NEWS) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death. *Resuscitation* 2013;84(4):465-70. doi:10.1016/j.resuscitation.2012.12.016

13. McPadden J, Durant TJ, Bunch DR, et al. Health Care and Precision Medicine Research: Analysis of a Scalable Data Science Platform. *J Med Internet Res* 2019;21(4):e13043. doi:10.2196/13043

14. Corey KM, Kashyap S, Lorenzi E, et al. Development and validation of machine learning models to identify high-risk surgical patients using automatically curated electronic health record data (Pythia): A retrospective, single-site study. *PLOS Medicine* 2018;15(11):e1002701. doi:10.1371/journal.pmed.1002701