User Guide and Documentation for the MIMIC II Database

Gari D. Clifford[1,2], Daniel J. Scott[1,2] and Mauricio Villarroel[1,2]

[1]Harvard-MIT Division of Health Sciences & Technology, Rm E25-505,
45 Carleton St., Cambridge MA 02142, USA
[2]Massachusetts Institute of Technology Technology, Rm E25-505,
77 Massachusetts Av., Cambridge MA 02139, USA

MIMIC-II database version 2.6.

Rev: 290 LastChangedDate: 2011-09-07 15:23:22 -0400 (Wed, 07 Sep 2011)

Preface

This user guide is intended for clinicians with some knowledge of programming and/or graduate-level researchers with knowledge of biomedical signal processing. The user is expected to have a working knowledge of `SQL`. Basic knowledge of a statistical (signal processing) package such as `Matlab` or **R** is useful. Knowledge of `C/C++` and/or `Java` may also help, but is not essential.

Many of the signal processing algorithms and data sets described in this guide are available from, or described in papers posted at the following URLs:

`http://www.physionet.org`
`http://mimic.physionet.org`

Most of the algorithms posted at the above URLs have been written either in C or `Matlab`. Libraries for reading these databases are also freely available. We hope that through these URLs this database will continue to evolve and add to the growing body of open (repeatable) biomedical research.

Gari Clifford, Li-wei Lehman, Roger Mark, Mohammed Saeed, George Moody, Daniel Scott, Ikaro Silva and Mauricio Villarroel

The Laboratory for Computational Physiology,
Cambridge, MA, USA,
January 2009

Acknowledgments

This work was supported by the National Institute of Biomedical Imaging and Bioengineering (NIBIB) of the National Institutes of Health (NIH) (grant number R01 EB001659). We also acknowledge contributions by Philips Healthcare and the National Library of Medicine. The content of this document is solely the responsibility of the authors and does not necessarily represent the official views of the NLM, the NIBIB, the NIH or Philips Medical Systems.

Of course, this database could never had come into existence without the hard work of all the students, staff and faculty. The MIMIC II database is a collective effort driven by several individuals guided by Professor Roger Mark from the Laboratory for Computational Physiology at the Massachusetts Institute of Technology. Many research collaborators were and are directly involved in the constant evolution of the database, including: Omar Abdala, Anton Aboukhalil, Tiffany Chen, Gari Clifford, Anagha Deshame, Margaret Douglass, Thomas Heldt, Isaac Henry, Caleb Hug, Brian Janz, Sherman Jia, Tin Kyaw, Li-Wei Lehman, Bill Long, Qiao Li, Christine Lieu, Atul Malhotra, Benjamin Moody, George Moody, Ishna Neamatullah, Tushar Parlikar, Andrew Reisner, Ali Saeed, Mohammed Saeed, Daniel Scott, Dewang Shavdia, James Sun, Peter Szolovits, Danny Talmor, George Verghese, Mauricio Villarroel and Wei Zong.

Several individuals made possible the data collection infrastructure, including Brian Gross, KP Lee, Larry Nielsen and Greg Raber from Philips Healthcare (Andover, MA) and Philips Research of North America, and John Halamka, Larry Markson, Larry Nathanson and Lu Shen from Harvard Medical School and the Beth Israel Deaconess Medical Center. We also would like to thank numerous other collaborators at MIT, Harvard, Beth Israel Deaconess Medical Center, Philips Research and our advisory committee: James B. Bassingthwaighte, Reed Gardner, Clement McDonald and Michael Shabot.

IRB Approval

This study was approved by the Institutional Review Boards of Beth Israel Deaconess Medical Center (Boston, MA) and the Massachusetts Institute of Technology (Cambridge, MA). Requirement for individual patient consent was waived as the study did not impact clinical care and all data were de-identified.

Document License

CHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU General Public License for more details.

You should have received a copy of the GNU General Public License along with this guide; if not, you can ask for a copy from:

Free Software Foundation, Inc.
675 Massachusetts Avenue
Cambridge, MA 02139
USA.

Comments and questions

We have revised the document to the best of our ability, but you may find that some features have changed since this document was published, that we made some mistakes, or you may simply need more information for a particular section. If so, please notify us by writing to:

MIMIC II Database
Laboratory for Computational Physiology
E25-505
Massachusetts Institute of Technology
77 Massachusetts Avenue
Cambridge, MA 02139
USA

Contents

Chapter 1

Introduction

1.1 Overview

This *User Guide* is intended to describe the MIMIC II (**M**ultiparameter **I**ntelligent **M**onitoring in **I**ntensive **C**are) database, an Intensive Care Unit (ICU) database which is freely available, together with this guide, from:

http://www.physionet.org/mimic2

This document can be viewed as HTML or downloaded as a PDF from the same location.

The MIMIC II database was collected as part of a Bioengineering Research Partnership (BRP) grant from the National Institute of Biomedical Imaging and Bioengineering entitled, "Integrating Data, Models and Reasoning in Intensive Care" (RO1-EB001659). The project was established in October 2003 and included an interdisciplinary team from academia (MIT), industry (Philips Medical Systems) and clinical medicine (Beth Israel Deaconess Medical Center). The objective of the BRP is to develop and evaluate advanced Intensive Care Unit (ICU) patient monitoring systems that will substantially improve the efficiency, accuracy and timeliness of clinical decision making in intensive care.

1.2 Background

ICU patients are typically the most physiologically fragile patients in the hospital and may experience prolonged hospital stays with significant morbidity and mortality. The modern ICU employs an impressive array of technologies that results in the generation of a rich–yet disparate–set of clinical and physiologic data used to guide patient care. ICU clinicians are challenged to interpret all the available ICU data to not only improve patient outcomes, but also to contain costs and adopt evidence-based practices. The enormous amount of ICU data and its poor organization make its integration and interpretation time-consuming and inefficient. The data overload that results may actually hinder the diagnostic process, and may even lead to neglect of relevant data, result-

ing in errors and complications in ICU care Tsien and Fackler [1997]. In the long term, automated or semi-automated monitoring and clinical decision support systems (CDSS) are needed. These systems must be capable of not only presenting ICU data to human users but also of forming pathophysiological hypotheses that best explain the rich and complex volume of relevant data from clinical observations, bedside monitors, mechanical ventilators and the wide variety of available laboratory tests and imaging studies. Such systems should reduce the ever-growing problem of information overload, and provide much more clinically relevant and timely alarms than today's disparate limit-based alarms. While there have been decades of research in utilizing artificial intelligence and expert-systems for medical data processing and forecasting Norris and Dawant [2002], little research has found its way into widely deployed ICU monitoring and information systems. The development of such systems requires access to large volumes of real-world ICU data that can serve as a testing platform to refine and evaluate such algorithms.

The role of a rich and comprehensive database in this context is twofold. First, through data mining, such a database allows for extensive epidemiological studies that link patient data to clinical practice and outcomes. Such insight can in turn motivate the development of alarms, alerts, or algorithms to improve clinical practice and thus improve patient outcomes. Second, it is essential to develop and test algorithms with real data, and to be able to perform such tests repeatedly and reproducibly as algorithm refinements evolve. Within the critical care community, well-known databases including APACHE Abbott et al. [1991] and Project IMPACT Glance et al. [2002] have resulted in the acquisition of hundreds of thousands of ICU patient cases from dozens of hospitals throughout the United States of America. The purpose of such databases is mostly to assess and compare the severity of ICU patient conditions and outcomes, and the costs of treatment across all participating intensive care units on the basis of very few, highly aggregate pieces of information. Such data abstractions often do not include detailed information regarding temporal relationships between therapeutic interventions and corresponding diagnostic data, and thus, would be insufficient to characterize clinically significant transient events such as hemodynamic instability, or acute organ injury.

The detail and volume of data necessary to support such research as described above has been difficult to gather in the past due to limitations on computational processing power, networking bandwidth, digital storage capacities, proprietary vendor data formats, and concerns related to patient privacy Clifford et al. [2009]. Through a collaborative effort between academia, industry, and clinical medicine, we have attempted to address these aforementioned challenges, and established a major new, publicly available ICU database, MIMIC II.

For more information on the rationale for assembling this database, the original research proposal can be found here:
http://mimic.physionet.org/BRP_Proposal.pdf

This document provides a detailed overview of the formation and contents of the MIMIC II database. The methodology used in data post-processing and the organization of MIMIC II is also described. In section 2 we provide a

characterization of MIMIC II with respect to quantitative data specifications as well as clinical characterizations using standard metrics such as patient acuities, problem lists, demographics, and mortality rates. In section 3 access modalities are described.

1.3 Overview of data collection

The data were collected over a seven year period, beginning in 2001, from Boston's Beth Israel Deaconess Medical Center (BIDMC). Any patient who was admitted to the ICU on more than one occasion may be represented by multiple patient visits. The adult ICUs (for patients aged 15 years and over) include medical (MICU), surgical (SICU), coronary (CCU), and cardiac surgery (CSRU) care units. Data were also collected from the neonatal ICU (NICU).

Figure 1.1 illustrates the data acquisition process, which did not interfere with the clinical care of patients, since databases were dumped off-line and bedside waveform data and derived trends were collected by an archiving agent over TCP/IP. Source data for the MIMIC II database consists of a) bedside monitor waveforms and associated numeric trends derived from the raw signals, b) clinical data derived from Philips' CareVue system, c) data from hospital electronic archives, and d) mortality data from the Social Security Death Index (SSDI). These data are assembled in a protected and encrypted database (both flat files for the waveforms and trends, and in the form of a relational database for all other data). Once the data have been assembled in a central repository and time aligned, the waveforms and trends for each individual are linked to the corresponding individuals' data in the relational database. (See section 1.4.3 for more information.) The data are then de-identified to produce a final set of data for public consumption. (See section 1.4.4 and Neamatullah et al. [2008] for more information on this detailed process.) We calculate standardized severity scores on the database and we also incorporate user feedback and corrections.

The resulting records contain realistic patient measures with all the associated challenges (such as noise or missing data gaps) that advanced monitoring and clinical decision support systems (CDSS) algorithms would receive as input data. Noise and artifact examples in the database, together with methods for dealing with these problems are described in sections 1.6 and 5.4.

1.4 Data Organization

1.4.1 Types of data

There are essentially two basic types of data in the MIMIC II database; clinical data stored in a relational database, and bedside monitor waveforms and his/her associated derived parameters and events stored in flat binary files (with ASCII header descriptors), and sorted with one directory per patient. Only a fraction of the total records in the relational database have associated waveform data. Of over 25,000 patients in the MIMIC II database, around 20,000 are adults

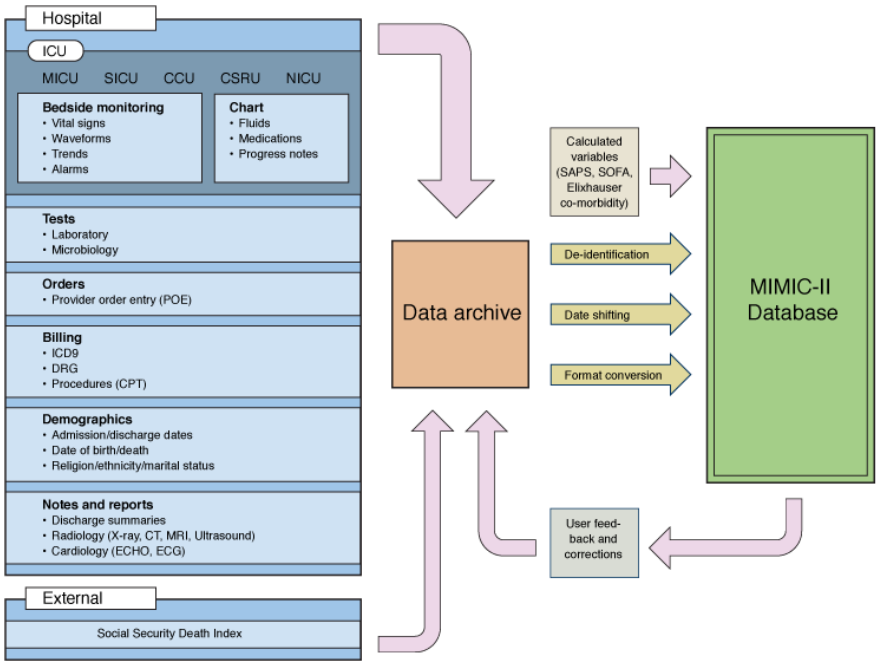


Figure 1.1: Schematic of data collection and database construction. Source data consists of: bedside monitor waveforms and trends, the ICU clinical databases, the hospital archives and the Social Security Death Index. These data are assembled in a protected and encrypted database which is then de-identified to provide one relational database plus associated flat file bedside waveforms and trends. We also incorporate standardized severity scores and user feedback and corrections.

and around 5000 are neonates. There are approximately 3000 waveform records of which around 2500 have been matched to the clinical data in the relational database.

Note that we define *waveforms* to be rapidly sampled (125Hz) signals recorded by the bedside monitors such as electrocardiograms (ECG) and arterial blood pressure (ABP) waveforms, illustrated in Figure 1.2. We define *trends* to be a time series of parameters derived from the waveforms by the bedside monitors, such as heart rate, systolic blood pressure, cardiac output and relative oxygen saturation. Of course, time series of repeated clinical measurements are also found in the relational database, such as pH levels, laboratory values and administered medications.

Figure 1.3 illustrates a typical set of time series (or 'trends'). The first two channels of data are HR (heart rate) and IBP (invasive blood pressure) which

Figure 1.2: Typical clean waveform data in the MIMIC II database. From top to bottom: Two leads of ECG (II and MCLI), arterial blood pressure (ABP) and pulmonary arterial pressure (PAP).

are taken from the flat file trend data. The third trace (NBP: non-invasive blood pressure) is recorded by a nurse from oscillometric cuff inflations and so is sampled much more sparsely. *Events* are automatically generated markers triggered by the bedside monitor algorithms. These include arrhythmia alarms, error messages (such as cable disconnections) and beat labels. These data are therefore unevenly sampled. Numeric trends are generally produced by the bedside monitors once per second, although after transmission to the central ICU database, they are often stored only once every 5 to 60 minutes. See section 2.3 for more details on these data types. A list of all the possible alarms can be found in table 2.6, ranked by their frequency in the database, together with associated statistics concerning the mean, minimum and maximum values at which the thresholds are set by the clinical staff.

Clinical data are recorded far less frequently than bedside monitor data and come from a variety of databases. These include the laboratory results, pharmacy provider order entry (POE) records, admission and death records,

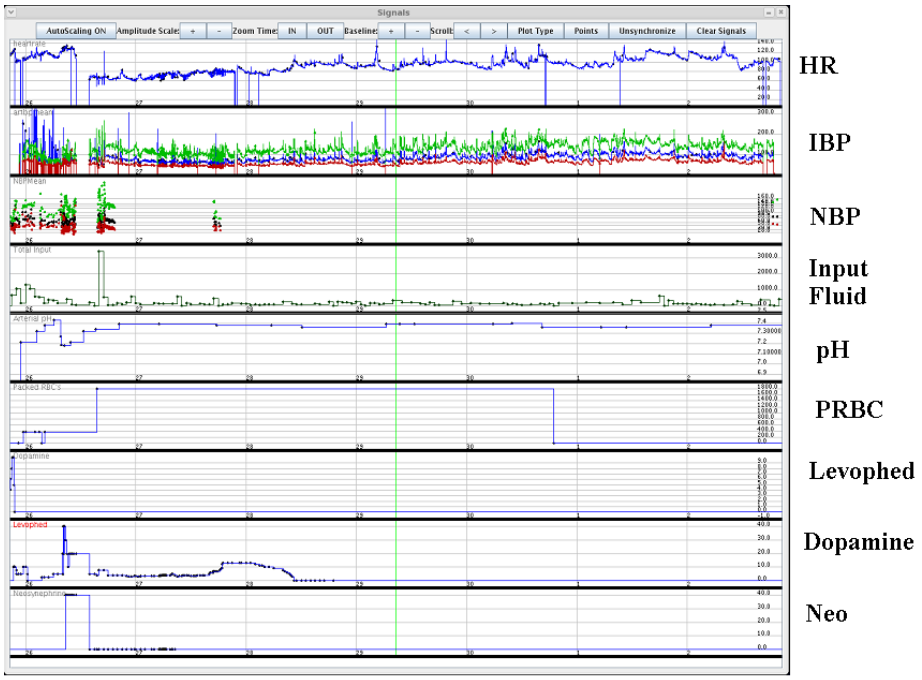


Figure 1.3: Trend data associated with a particular patient stay. Parameters are HR: heart rate, IBP: invasive blood pressure (systolic, mean and diastolic in green, blue and red respectively), NBP: non-invasive blood pressure (with the same color coding), Input Fluid: total fluids given to the patient per hour, pH: acidity/alkalinity of patient, PRBC: packed red blood cell administration, Levophed: Levophed administration, Dopamin: dopamine levels, and Neo: neosynephrine.

demographic details, discharge summaries, ICD-9 codes, procedure codes, microbiology and lab tests, imaging and ECG reports and the ICU central database (which includes some subset of the bedside monitor trends, drip rates, free text nursing notes and nurse-verified down-sampled trends, amongst other information). A selection of these parameters can be seen in Figure 1.3. A more detailed description of the content of these databases can be found in section 2.1.

1.4.2 What is a patient record?

Since a patient may have been admitted several times during the period in which our data were collected, it is important to understand exactly how to identify patients and his/her stay(s).

There are essentially four identifiers for data associated with any given patient:

- Subject ID (*Subject_ID*) - an integer number identifying a particular patient. This can be thought of as a substitute for a unique medical record number. In the flat file data posted on PhysioNet, the number representing the Subject_ID is left padded with zeros to five digits and preceded by the letter *s*. In the relational database, the Subject_ID has no preceding letter or leading zeros.
- Hospital admission ID (*Hadm_ID*) - an integer number identifying a particular admission to the hospital. Each patient may have many *Hadm_ID*s associated with his/her unique *Subject_ID*.
- ICU stay ID (*ICUstay_ID*) - an integer number identifying an ICU stay. An ICU stay, refers to the period of time when the patient is cared for continuosly in an Intensive Care Unit. Each patient may have one or more ICU stays associated. An ICU stay is considered to be continuous if any set of ICU events (such as bed transfers or changes in type of service) belonging to one Subject_ID which are fewer than 24 hours apart. Longer breaks in the patient's stay automatically cause a new *ICUstay_ID* to be assigned.
- Case ID (*Case_ID*) - This is a five digit number preceded by the letter *a* (for adults) or *n* (for neonates). This ID indicates a set of waveforms associated with a given patient. For various reasons (described in section: 1.4.3 below), there may be multiple case IDs associated with a given patient.

Figure 1.4 illustrates the possible data available for a given individual, identified by a "subject_id". Time progresses from left to right, and the different types of data collected are shown vertically. Each subject can have multiple hospital admissions, identified with "hadm_ids". Each hosptial admission can contain multiple ICU stays, identified with "icustay_ids". Waveforms collected during ICU stays are identified using "case_ids". Laboratory and microbiology tests are performed throughout a hospital stay and can therefore take place outside the ICU stay. Vital sign validation, medications, fluid balances and nursing notes are only performed in the ICU and are not available during the remainder of the hospital stay. Date of death is recorded in-hospital and has also been obtained from social security records for out-of-hospital mortality.

The above illustrates an "ideal" case where the timestamps associated with the data fall within the hospital and/or ICU stay. Unfortunately, real-world issues can complicate matters allowing data to be recorded outside of a patient stay. For example, a patient could be physically present in the ICU and connected to monitors before their admission has been entered into the system. This results in a waveform recording which starts before the subject's ICU admission. Furthermore, missing/mistaken data can mean that ICU stays exist where there is no matching hospital admission record.

Note that a patient may move between ICUs during any given admission. If the move is longer than 24 hours, we define it to be a new ICU stay. Note also that the amount of data varies during and between ICU stays and that data are often missing - see section 1.6.

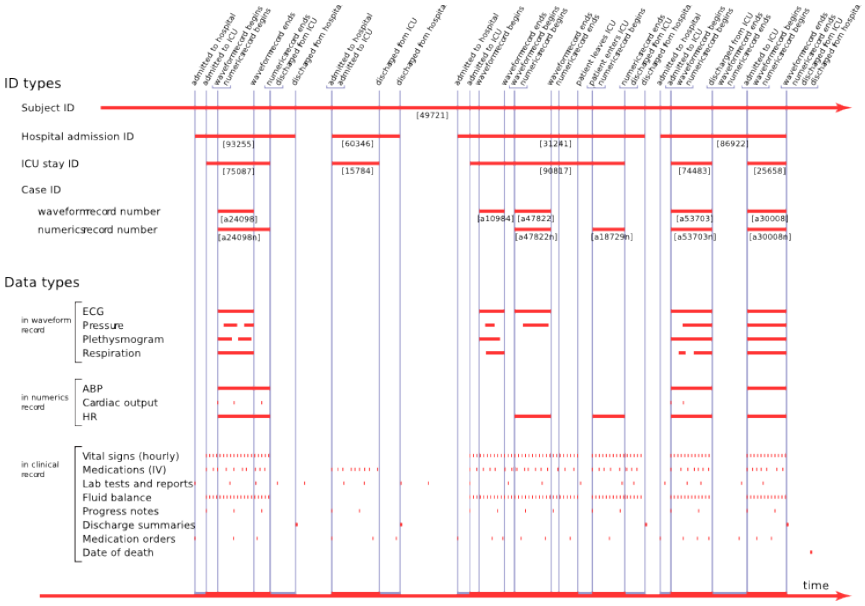


Figure 1.4: Schematic of a patient record. Note that the patient may experience several hospital admissions and ICU stays, for which differing amounts of data are available. See section 1.4.2 for details.

1.4.3 Subject ID - Case ID matching

Given that the MIMIC II data are collected from different sources, they must be matched to a unique patient and temporally aligned. The bedside monitor-generated data included a unique identifier (the *Case_ID*), assigned automatically by the monitor, and fields for patient name (first and last name) and medical record number (MRN). The name and MRN fields were manually entered by nurses into the networked central station when a patient was admitted. Unfortunately in approximately 30% of cases one or more identifier fields were not completed for admitted patients. Moreover, human errors are likely to exist in the manually recorded name and MRNs. The CareVue clinical information system also included a unique patient identifier (that maps to our *ICUstay_ID*) for each ICU stay of a patient. The subject's CareVue data also includes identifying information such as a patient's name and MRN which was automatically input from the hospital-wide information system when a patient is admitted to a unit.

When waveform files included the patient's identifying information (name, and MRN), the physiologic data records (indexed by a *Case_ID*) were matched to the corresponding clinical information records from CareVue. There were two stages to the merging process. The first stage included matching names and medical record numbers (when available and accurately recorded) from

the monitor-generated data records to those of the clinical data records from CareVue. The second stage included comparing the similarity of the physiologic trends from the higher resolution monitoring data (approximately 1 sample per minute) with the nurse-validated vital sign trends in the clinical information system sampled on an hourly basis.

Briefly, determination of trend similarity included four stages:

- Determining a temporal overlap between the available trends that were present in the physiologic and clinical data.
- Identifying if unusual parameters had been recorded at the same time (e.g. cardiac output).
- Correlating median filtered, down-sampled heart rate and blood pressure numeric trends with heart rate and blood pressure trends in the clinical data.
- Visual inspection of a subset of files to verify the results.

However, it is possible that some of the matches may be incorrect. After manual review we believe we have caught most of the inconsistencies, but anomalies may still be present.

1.4.4 De-identification of patients' data

The process for the removal of protected health information (PHI) in the the MIMIC II database is fully described in Neamatullah *et-al*Neamatullah et al. [2008]. A labeled subset of the data, together with a public version of the code can be found on PhysioNet at: `http://www.physionet.org/physiotools/deid/`.

Figure 1.5 illustrates the de-identification process. Briefly, the salient points for the user of our database are:

- All dates were shifted into the future.
- All ICU dates for a given patient were shifted by the same amount to preserve inter-admission time gaps.
- The day of the week and season of the year were preserved.
- Patients who turned 90 during one of his/her admissions have been removed from the database. They may be included at a later date.
- Patients older than 89 years at the date of first admission have had his/her dates of birth shifted so that they appear to be 200 years old at the time of his/her first admission. They will therefore show up as extreme outliers. The inter-admission timings are still preserved.

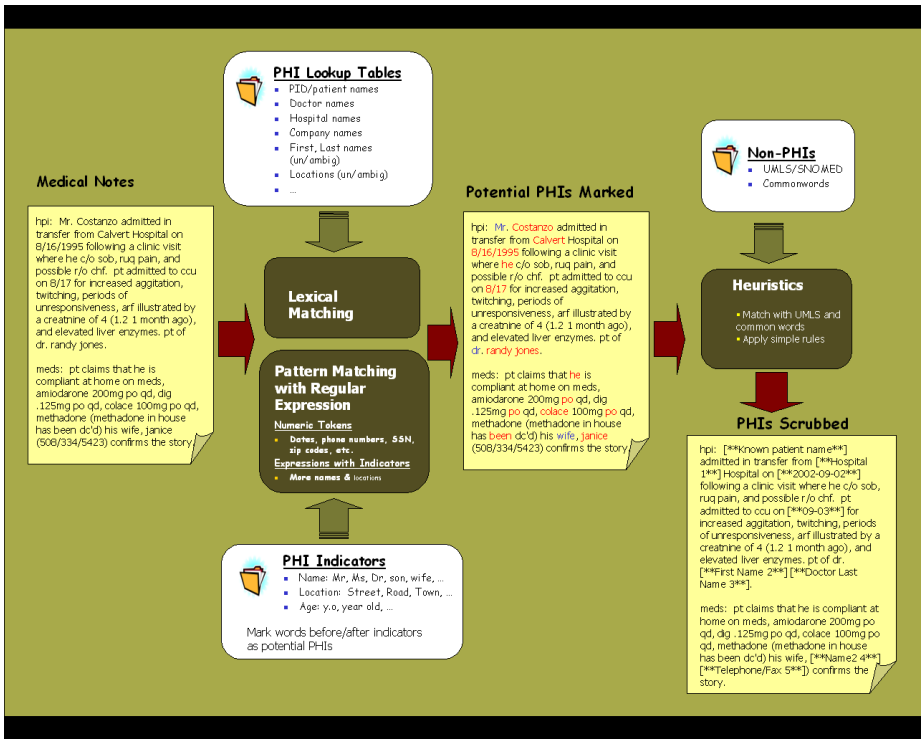


Figure 1.5: De-identification process

- Since date shifts were randomly assigned, longitudinal studies that involve changes in patient care practices over time cannot be supported by the fully de-identified data. Support for studies that require the year of admission will be considered on an individual basis by special request.

- All HIPAA-defined types of PHI were removed, plus care-giver and hospital-specific identifiers.

- The algorithm achieved an overall recall of 0.967 and precision of 0.749 on a 'gold standard' test corpus, which out-performs a single human de-identifier and performs at least as well as a consensus of two human de-identifiers Neamatullah et al. [2008].

Examples of a de-identified nursing progress note and discharge summary can be found in figures 1.6 and 1.7 respectively. Note that a few of the de-identified sections of the nursing note are false positives, and a small fraction of the clinical information may have been lost. However, all dates and names (the only PHI in this document) were caught by our algorithm. Note also the the high prevalence of abbreviations such as S/O (sign out), D/C'd (discontinued,

or discharged), Neo (neosynephrine), NSR (normal sinus rhythm), F/E (fluid and electrolytes), GI (gastrointestinal), HEME (hematology), ID (infectious disease), A (assessment), P (plan), etc. Note also the low degree of structure in the nursing note, broken into a few categories; **S/O, F/E, NEURO, GI, HEME, ID, RESP, SKIN, ACCESS, SOCIAL, A,** and **P**. The boldface type has been added to this figure to highlight these categories, but is not available in the notes.

1.5 Clinical overview of patients in the initial release of the MIMIC II database

As of version 2.6 (April 2011) MIMIC II contains around 33,000 patients of which approximately 25,000 are adults (having age $\geq$ 15 years old at time of last admission) and around 8000 are neonates (age $\leq$ 1 month old at the time of first admission. These patients experienced over 36,000 hospital admissions and over 40,000 ICU stays.

A statistical summary of the database contents is provided for each release of the database and can be found on the project's website:

https://mimic.physionet.org/database/releases.html

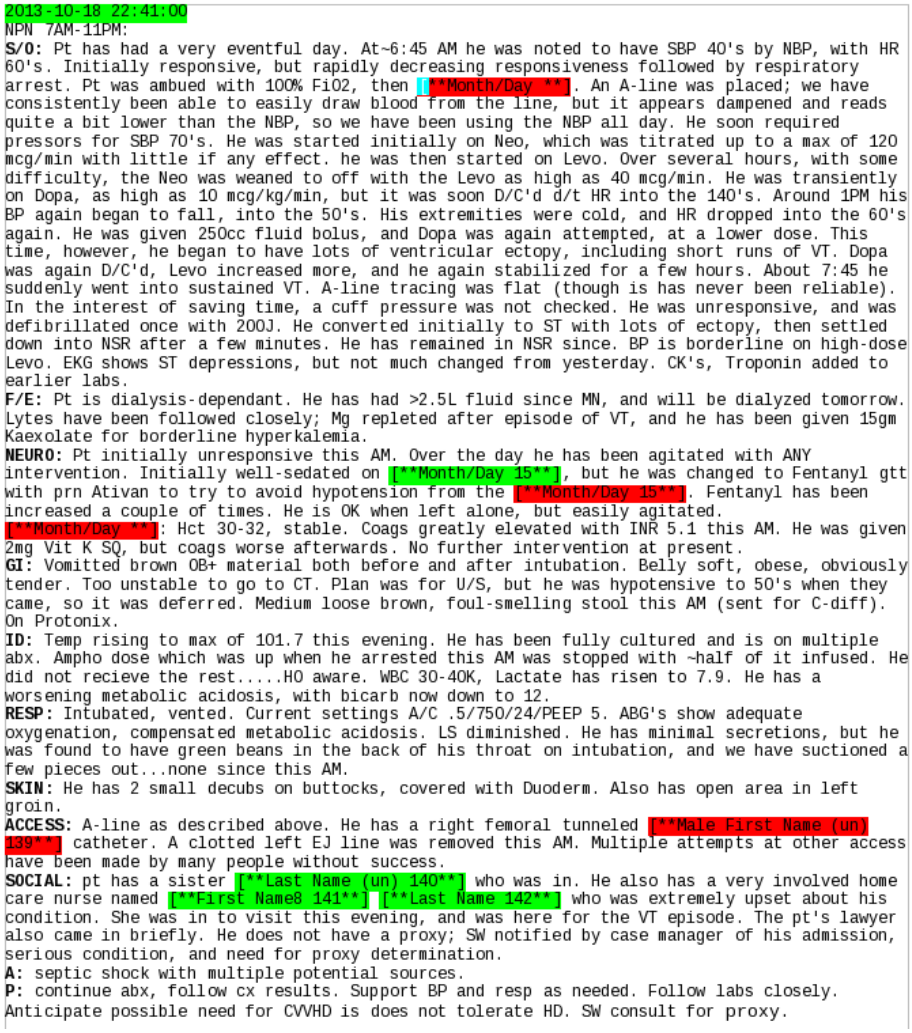
Although we have collected neonatal waveforms, these will likely be released at a later date when the associated clinical data has been matched and verified. The rest of the description of the data in this section is therefore limited to the adult population.

Figure 1.8 illustrates the age distribution of adult patients entering the ICU. Therefore, a patient will be represented more than once if they were admitted to the ICU more than once. The mean age at time of admission was 63.44 years and the median age was 65.33 years. Note small drops in admission rates for patients during his/her mid 20s to mid 30s, and again around 65 years of age. The distribution of length of stay for each ICU admission is illustrated in figure 1.9. Note that the distribution is heavy tailed, with a mean of 4.57 days and a median of only 2.15 days.

1.6 Noise, artifacts and missing data

The data we have collected is highly representative of that which can be found in the ICU, and therefore is replete with noise and artifacts due to patient movement, sensor degradation, transmission errors, electromagnetic interference and human error.

We have begun to develop signal quality indices to label useful and noisy sections of the data. Currently we have signal quality indices for the ECG and blood pressure waveforms, and we are close to completion for the pulse oximeter and the respiratory waveforms. A more detailed description of the algorithms for signal quality can be found in section 2.3.5 and in Zong *et al.* Zong et al. [2004], Sun *et al.* Sun et al. [2006] and in Li *et al.* Li et al. [2008, 2009].



2013-10-18 22:41:00
NPN 7AM-11PM:
S/O: Pt has had a very eventful day. At~6:45 AM he was noted to have SBP 40's by NBP, with HR 60's. Initially responsive, but rapidly decreasing responsiveness followed by respiratory arrest. Pt was ambued with 100% FiO2, then [**Month/Day **]. An A-line was placed; we have consistently been able to easily draw blood from the line, but it appears dampened and reads quite a bit lower than the NBP, so we have been using the NBP all day. He soon required pressors for SBP 70's. He was started initially on Neo, which was titrated up to a max of 120 mcg/min with little if any effect. he was then started on Levo. Over several hours, with some difficulty, the Neo was weaned to off with the Levo as high as 40 mcg/min. He was transiently on Dopa, as high as 10 mcg/kg/min, but it was soon D/C'd d/t HR into the 140's. Around 1PM his BP again began to fall, into the 50's. His extremities were cold, and HR dropped into the 60's again. He was given 250cc fluid bolus, and Dopa was again attempted, at a lower dose. This time, however, he began to have lots of ventricular ectopy, including short runs of VT. Dopa was again D/C'd, Levo increased more, and he again stabilized for a few hours. About 7:45 he suddenly went into sustained VT. A-line tracing was flat (though is has never been reliable). In the interest of saving time, a cuff pressure was not checked. He was unresponsive, and was defibrillated once with 200J. He converted initially to ST with lots of ectopy, then settled down into NSR after a few minutes. He has remained in NSR since. BP is borderline on high-dose Levo. EKG shows ST depressions, but not much changed from yesterday. CK's, Troponin added to earlier labs.
F/E: Pt is dialysis-dependant. He has had >2.5L fluid since MN, and will be dialyzed tomorrow. Lytes have been followed closely; Mg repleted after episode of VT, and he has been given 15gm Kaexolate for borderline hyperkalemia.
NEURO: Pt initially unresponsive this AM. Over the day he has been agitated with ANY intervention. Initially well-sedated on [**Month/Day 15**], but he was changed to Fentanyl gtt with prn Ativan to try to avoid hypotension from the [**Month/Day 15**]. Fentanyl has been increased a couple of times. He is OK when left alone, but easily agitated.
[**Month/Day **]: Hct 30-32, stable. Coags greatly elevated with INR 5.1 this AM. He was given 2mg Vit K SQ, but coags worse afterwards. No further intervention at present.
GI: Vomitted brown OB+ material both before and after intubation. Belly soft, obese, obviously tender. Too unstable to go to CT. Plan was for U/S, but he was hypotensive to 50's when they came, so it was deferred. Medium loose brown, foul-smelling stool this AM (sent for C-diff). On Protonix.
ID: Temp rising to max of 101.7 this evening. He has been fully cultured and is on multiple abx. Ampho dose which was up when he arrested this AM was stopped with ~half of it infused. He did not recieve the rest.....HO aware. WBC 30-40K, Lactate has risen to 7.9. He has a worsening metabolic acidosis, with bicarb now down to 12.
RESP: Intubated, vented. Current settings A/C .5/750/24/PEEP 5. ABG's show adequate oxygenation, compensated metabolic acidosis. LS diminished. He has minimal secretions, but he was found to have green beans in the back of his throat on intubation, and we have suctioned a few pieces out...none since this AM.
SKIN: He has 2 small decubs on buttocks, covered with Duoderm. Also has open area in left groin.
ACCESS: A-line as described above. He has a right femoral tunneled [**Male First Name (un) 139**] catheter. A clotted left EJ line was removed this AM. Multiple attempts at other access have been made by many people without success.
SOCIAL: pt has a sister [**Last Name (un) 140**] who was in. He also has a very involved home care nurse named [**First Name8 141**] [**Last Name 142**] who was extremely upset about his condition. She was in to visit this evening, and was here for the VT episode. The pt's lawyer also came in briefly. He does not have a proxy; SW notified by case manager of his admission, serious condition, and need for proxy determination.
A: septic shock with multiple potential sources.
P: continue abx, follow cx results. Support BP and resp as needed. Follow labs closely. Anticipate possible need for CVVHD is does not tolerate HD. SW consult for proxy.

Figure 1.6: Example of a de-identified progress note. Sub-headings have been capitalized in bold face type for easier reading. Removed text is denoted by square brackets. True positives are colored green, false positives are colored red.

==========

DISCHARGE SUMMARY

Name: [**Known patient lastname**], [**Known patient firstname**]

[**Unit Number 626**]

Admission Date: [**2016-11-07**]

Discharge Date: [**2016-11-22**]

Date of Birth: [**1972-09-20**]

Sex: F

HISTORY OF PRESENT ILLNESS: Patient is a 44-year-old lady status post living related kidney transplant on [**2016-10-19**], who presented at [**Hospital 36**] for end-stage renal disease secondary to type 1 diabetes mellitus.

She presented to [**Hospital1 **] on [**2016-11-07**] with increased drainage from her surgical wound and JP, increased abdominal pain, and anuria x4 days. The patient reported constipation for a week. She denies flatus. She was complaining of nausea and vomiting. Her abdominal pain had become progressively worse left lower quadrant most notable. There is no radiation to the back or elsewhere. She denied any fevers, chills. She noted decreased p.o. intake recently. Her drainage from her wound incision and JP was notable for yellowish clear urine smelling fluid.

==========

Figure 1.7: Example of a section of a de-identified discharge summary. All de-identified elements are denoted by square brackets. No false positives exist in this example.

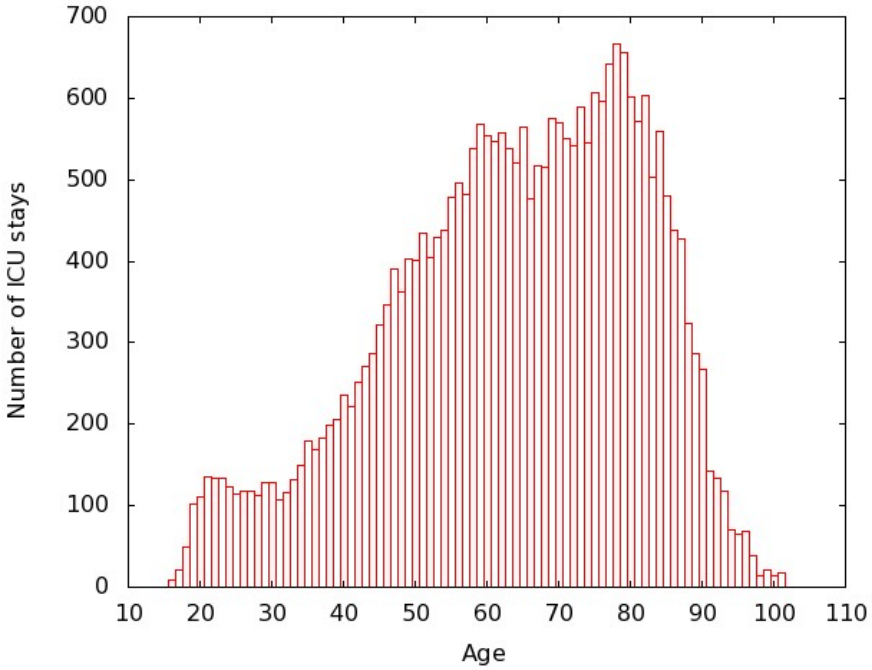


Figure 1.8: Age distribution of adult patients at time of ICU admission in MIMIC II database.

Signal quality indices for trend data are almost impossible to generate, except by using thresholds on gradients and absolute values that are physiologically impossible. Generally it is better to refer back to the original underlying waveform to derive a signal quality metric.

Data are also missing due to machine or patient disconnections, transmission and recording errors, or human omissions. Some data are also not requested very frequently, and so, although not technically missing, important events may go unobserved.

Although short-term missing data can be mitigated somewhat through interpolation, much of our non-waveform and trend data is sparsely sampled. Moreover, the data is not missing at random, since it can be due to changes in shifts or staff-to-patient ratios, or simply because a clinician or nurse did not think that the data were important. Interpolation, or imputation is therefore impossible, unless a model of how the data are missing can be constructed.

Apart from these problems, there may also be errors in the data matching and alignment, particularly where the data come from different sources and use different clocks. The clocks are not always rigorously synchronized and may drift. The problem is particularly apparent during the beginning/end of daylight saving time. Section 5.4 details these and other known issues with the

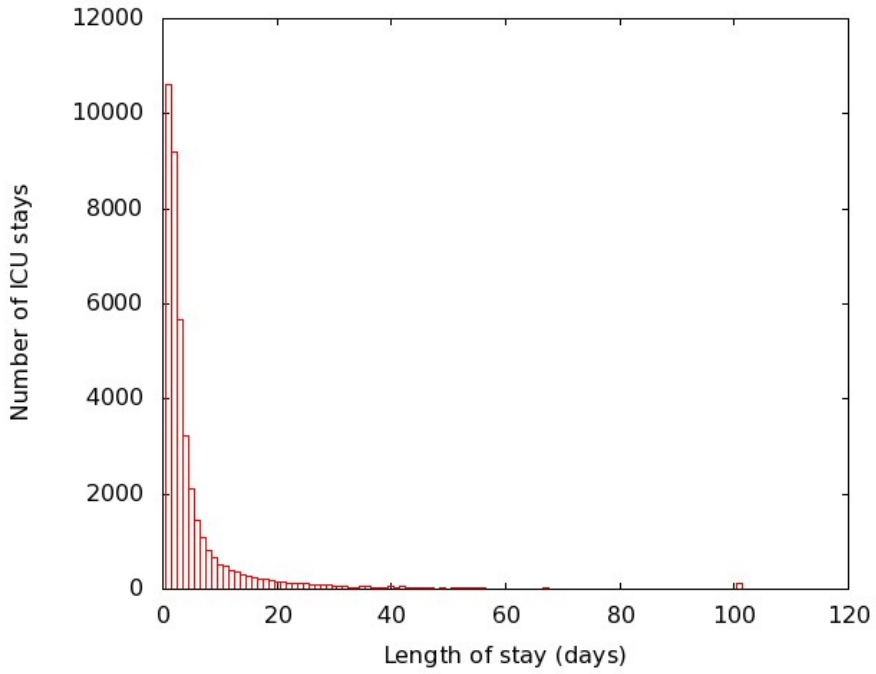


Figure 1.9: Length of stay of patients in MIMIC II database.

data.

1.7 Summary and further reading

Details of the issues surrounding data collection and annotation can be found in Clifford et al. [2009] Aboukhalil et al. [2008] and Saeed et al. [2009], and details related to de-identification can be found in Neamatullah et al. [2008]. Although the database described in this document is large and detailed, caution should be taken when analyzing the data, since the clinical systems for monitoring and archiving the data are not perfect and some data may be incorrect.

Moreover, we were unable to capture all of the data associated with a patient's stay, since we did not have data collection facilities in the OR, ER, or step-down wards. Neither were we able to obtain data from other hospitals for any given individual, so visits to outpatient care, or to other hospitals will not be reflected in a patient's stay. Finally, no patient data before the start of the project is available, other than the pre-existing history recorded at a particular stay. Despite these limitations, there is a wealth of data to be found in MIMIC II.

Note that this document does not provide detailed information on how to extract data from the WFDB flat-file versions of the data found in the MIMIC II database. More information on this can be found here:

http://www.physionet.org/physiobank/tutorials/using-mimic2/

and here:

http://www.physionet.org/physiotools/getting-started.shtml

Chapter 2

Database Description

2.1 Overview

The MIMIC II database is composed of two distinctive groups of data. The first group, the clinical database, consists of data integrated from different information systems in the hospital and contains diverse information such as: patient demographics, medications, results of lab tests and more. The second group, contains high resolution waveforms recorded from the bedside monitors in the intensive care units.

2.2 Clinical database

The MIMIC II clinical database, is a relational database. Although not strictly required, some familiarity with the structured query language (SQL) will help to understand how the data is stored and obtained. The purpose of SQL is to provide an easy interface to a given database. In relational databases, a typical SQL sentence has the following structure:

```
SELECT column_list
FROM datasources
WHERE constraints
```

Where:

- The SELECT statement, specifies which columns should be returned.
- The FROM statement, denotes the name of the tables where the data is stored.
- The WHERE clause, filters the data retrieved by some criteria.

The query in listing 2.1, shows a real case example: "Extract basic information for a subset of patients". The result of the query is shown in Table 2.1. Although this query is relatively simple, most of the queries throughout this document follow the same structure. The queries presented in the following sections were written with the goal to be illustrative and are not optimized in any way. Detailed explanation of SQL language and query optimizations are beyond the scope of this document, but can be easily be found in SQL Beaulieu [2005].

Listing 2.1: Extracting basic information about patients.

```
SELECT subject_id, sex, dob
  FROM d_patients
 WHERE subject_id in (7049, 7060, 7072,
       7078, 9181, 9185, 9195)
```

Figure 2.1 summarizes the major database components, their corresponding attributes and how they relate with a particular patient. The full database schema is much more complex than the one displayed in Figure 2.1, a full description of each table, along with column data types, naming convention, indices, relationships and constraints are provided online (http://mimic.physionet.org/schema/latest).

In the following sections, we provide a general overview, the list of database tables involved, and some sample data for each major group.

2.2.1 Patient

A patient in the MIMIC II database is uniquely identified by an integer number called *Subject_ID*, it can be thought of as a medical record number (MRN) normally found in hospital information systems. The basic information for any given patient is stored in the table D_PATIENTS, normally referred as the patient table throughout this document. As the database went through a careful de-identification process, the patient table only stores the patient identifier (*Subject_ID*), gender (sex) and date of birth (dob, shifted). Table 2.1 shows a sample content of the patient table resulted from the query in listing 2.1.

The date of death for patients who died in the hospital is taken to be the date of discharge. For other patients, date of death was obtained from social security death records from the US government. Where there is a conflict between the hospital data and social security, the hospital data is assumed to be more accurate. Note that patients who left the US (and subsequently died) after receiving treatment may not have their date of death recorded in the social security archives.

As shown in Figure 2.1, the patient identifier (*Subject_ID*) is widely used by most of the tables throughout the database to specify to which patient a given measurement or recording refers to. Figure 2.2 shows an example relating which diagnosis codes (ICD-9) were assigned to a given patient, the *Subject_ID*

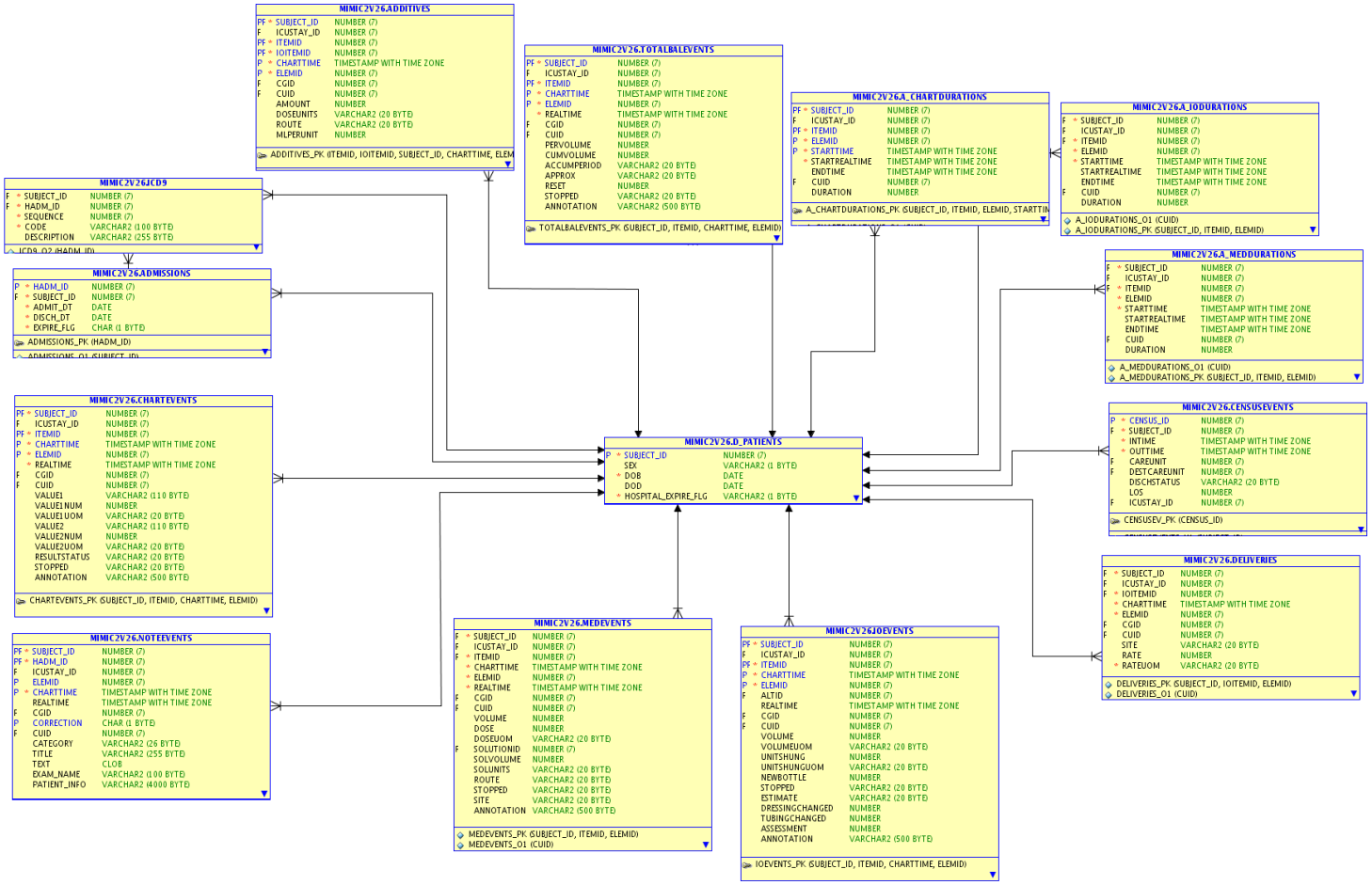


Figure 2.1: Major MIMIC II clinical database components. The patient table D_PATIENTS is central to the database model, other information such as admission, dempgraphics, procedures or medications can be readily accessed once a particular patient is identified. The tables above show the relationship between the major components in the database.

SUBJECT_ID	SEX	DOB	DOD	HOSPITAL_EXPIRE_FLG
7049	M	04/10/1952	03/18/2020	N
7060	F	08/01/1932	(null)	N
7072	M	02/22/1928	03/20/1999	Y
7078	F	11/11/1967	10/17/2012	Y
9181	F	03/11/1960	02/16/2007	Y
9185	M	02/28/1927	(null)	N
9195	F	12/19/1974	(null)	N

Table 2.1: Sample content of the patient table. Each patient has a *Subject_ID*, gender date of birth, date of death and flag indicating whether or not the patient died in the hospital.

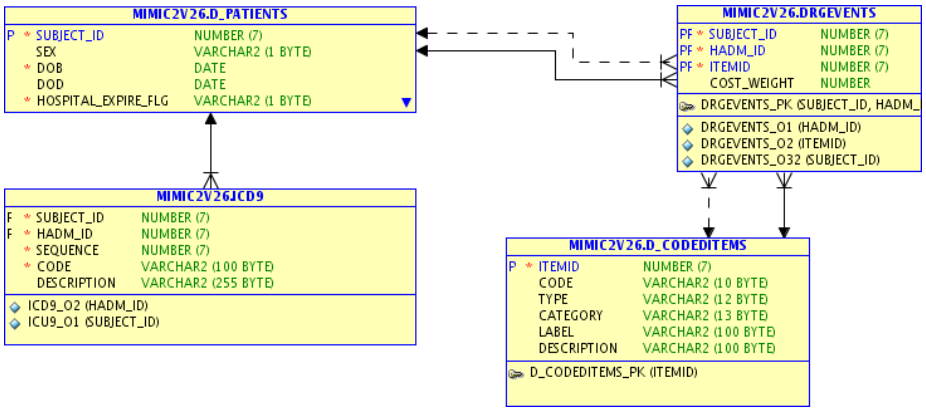


Figure 2.2: Patient to ICD-9 and diagnosis-related group code relationships. The patient's ICD-9 codes are related to the patient record in the patient table via a foreign key on the *Subject_ID* field and to the hospital admission in the admissions table through the hadm_id. DRGs, found in the drgevents table are related in the same way. The full meanings of the DRG codes are stored in d_codeditems.

field links the ICD-9 and the patient tables. The ICD-9 table records the ICD-9 codes applied to a particular patient during a specific hospitalization period. Diagnosis-related groups (DRGs) are stored in the drgevents table and their meanings are stored in d_codeditems.

2.2.2 Care Giver

Caregivers are the medical staff who are responsible for patients during their hospital stay such as: nurses, residents or other clinicians. The caregivers are

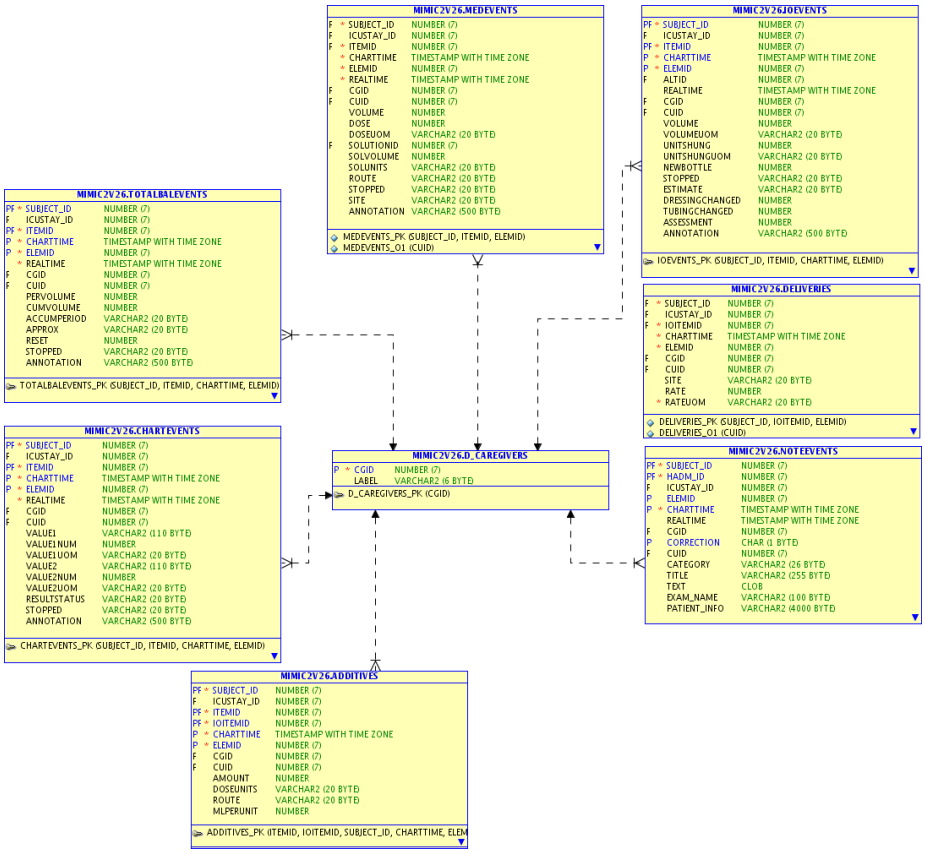


Figure 2.3: Caregivers table relationships. Although a simple table with only 2 columns, the caregivers table is related to many other tables in the database. Caregivers are assigned to *inter alia* medical events, problems and notes.

stored in the D_CAREGIVER table, and are uniquely identified by their care giver id (cgid). The caregivers table is related to many other tables such as medevents, noteevents and chartevents and is used to record the care giver who performed a particular operation, procedure or event. Figure 2.3 shows the inter table relationships of the caregivers table.

2.2.3 Care Unit

The Careunits table, stores information pertaining to the different ICU rooms in the hospital. Figure 2.4 shows the careunit table and its relationship with other event tables in the database. When a note is filed, a problem occurs or a chart event is entered, the particular care unit in which the event took place is recorded. The careunit table only contains a cuid (care unit id) and the unit

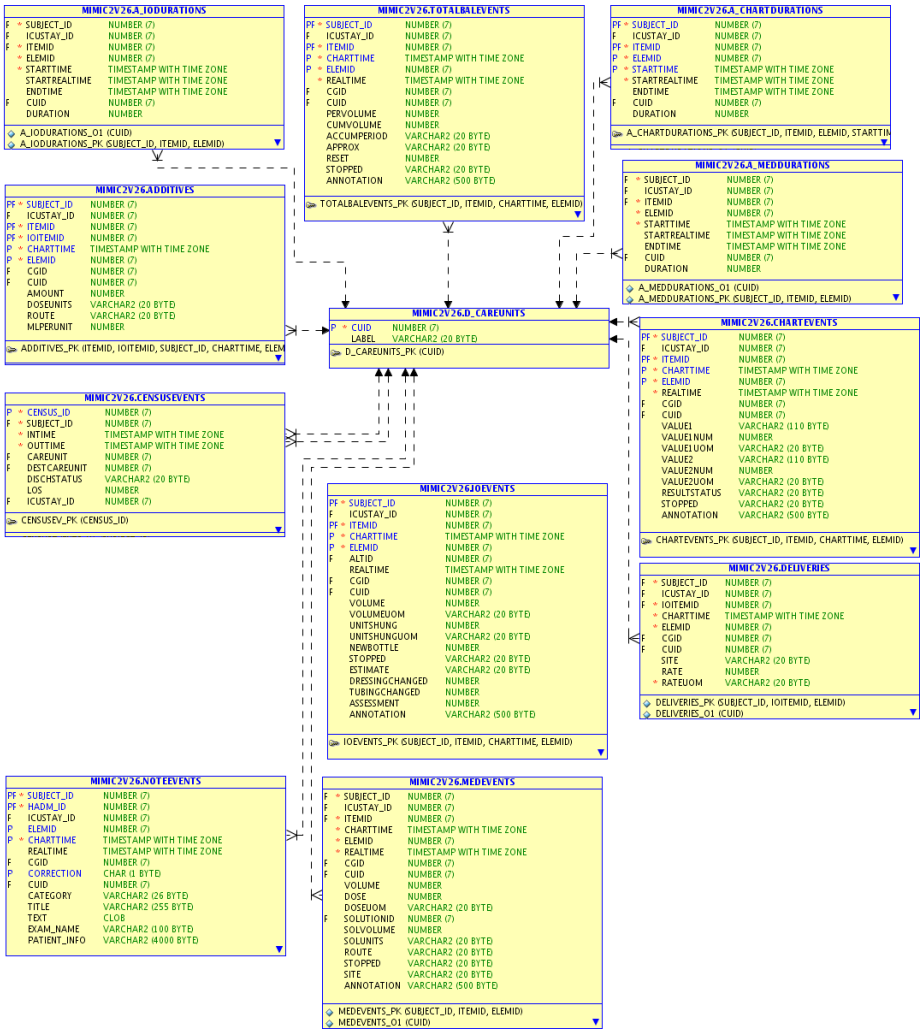


Figure 2.4: Careunits table and relationships. The careunits table is related to many other tables and contains information about the particular care unit where an event occurred.

name, but is used in many places to record the location in which events took place. Table 2.2 shows some sample content.

2.2.4 Patient timeline

When patients arrive at the hospital, and during their course of their stay, they can be transferred between different units, sent to the operating room for

UNIT NAME
CCU
CSRU
Nursery
Regular ward
Labor and Delivery
Outside hospital Cardiac ICU

Table 2.2: Sample content of the care units table.

surgery, sent to the floor for recovering or undergo other procedures. To better describe these events, Figure 2.5 shows an excerpt from a discharge summary for a typical patient.

Figure 2.6 is the visual representation of the events presented in Figure 2.5. We can identify the following events:

Hospital admission :

A hospital admission, covers the period from the patient's admission to the hospital, until the patient's discharge from the hospital. It includes any visits to different wards (such emergency room, regular floor, and even different stays in an ICU room).

Patient admissions are recorded in the admissions table. As well as recording the *Subject_ID* of the admitted patient, each admission has a unique identifier (Hadm_ID) and an admission and discharge time.

Figure 2.7) shows the relationship between the admission table and other tables in the database.

ICU Stay :

An ICU stay is a combination of one or more ICU census events that are separated by 24 hours or less.

ICU census event :

Each time a patient enters or leaves a particular care unit, an event is recorded into the database in the table censusevents. Each of these events (identified by the column census_id), contain the time and date of entrance and exit of the care unit, the current unit the patient was hosted, the destination care unit the patient was transferred to, and the length of stay in the ICU room for that particular event.

2.2.5 Patient data

During a patient's hospital stay, various information is collected about a patient. Demographics, vital signs, laboratory tests, medications, fluid balance, nursing notes, imaging reports, etc. can all be recorded in the database.

...Chief Complaint:
74 year old female admit to [**Hospital1 80**] [**Hospital Unit Name 26**] [**2018-05-16**] in resp distress, pna, UTI, mild CHF initially on NRB, but then intubated on [**2018-05-18**] (extubated [**2018-05-23**]). Hosp course noted for bradycardia (AV block) during swan placement, CHF. PMH: recent MI, CHF, a fib, CVA, GERD, gastritis, TIAs, Bell's palsy, lower GI polyps...

* * *

...transfered from outside hospital status post embolecotomy for R brachial emboli with history of severe aortic stenosis and anemia for cardiac work-up. Patient of Dr. [**Last Name (STitle) **], found to have colonic polyps on colonoscopy for anemia work-up at OSH. Pt admitted to receive medical clearance for future procedure. Found to have a urinary tract infection with signs of sepsis severe respiratory difficulty and severe aortic stenosis...

* * *

...On [**2018-06-27**], Ms. [**lastname 6384**] was taken to the operating room where she underwent an aortic valve replacement utilizing a 21mm [**Last Name (un) **] [**Doctor Last Name **] pericardial bioprosthesis. Postoperatively she was taken to the cardiac surgical intensive care unit for monitoring....

* * *

...She developed atrial fibrillation and underwent cardioversion on [**2018-06-30**]. Ms. [**lastname 6384**] was only able to hold a normal sinus rhythm for less then two minutes and amiodarone was started. Heparin and coumadin were started for anticoagulation with the plan for a repeat cardioversion in a month. Tube feeds were started for nutritional support and calorie counts were started. On postoperative day seven, Ms. [**lastname 6384**] was transferred to the cardiac

Figure 2.5: An excerpt from a patient's discharge summary, describing typical events during the patient stay in the hospital.

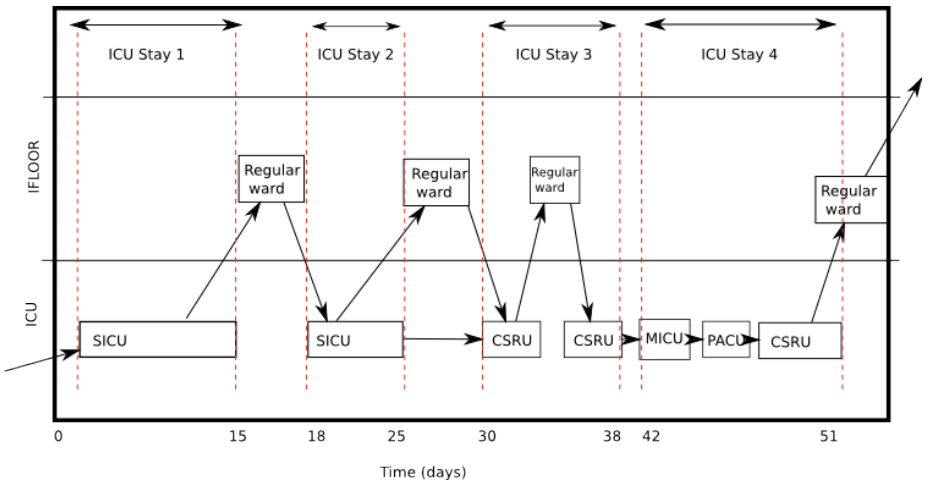


Figure 2.6: Typical events during the patient hospitalization period.

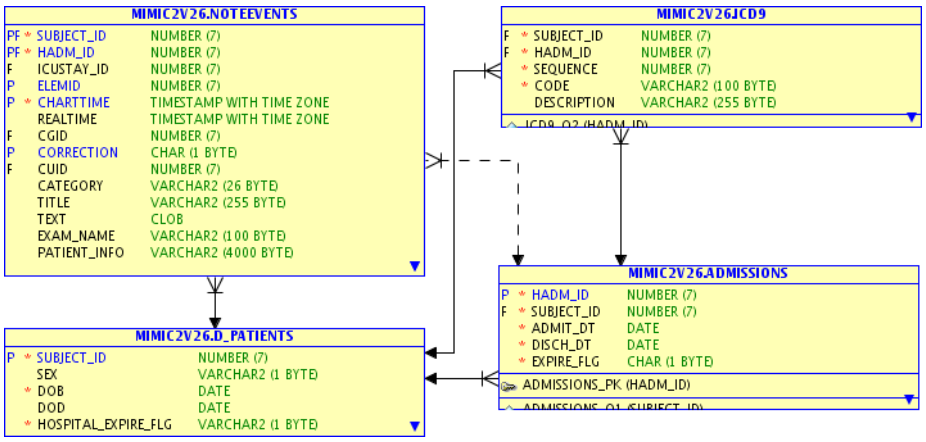


Figure 2.7: Relationship between a hospital admission and other database tables.

- Demographics: When a patient is admitted to the hospital, their demographic information is recorded. Marital status, ethnic origin, religion and insurance status are recorded.
- Medication records: Medications prescribed and administered either via computer controlled iv or oral. Computer controlled administration via iv is automatically administered. Although manual prescription is recorded for pharmacy orders, there is no guarantee that the medication was actually administered to the patient.
- Fluid records: Fluids withdrawn/administered from/to a patient are also recorded. This provides physicians with an up-to-date and accurate measure of a patient's fluid levels.
- Notes: MD Notes (for neonates), nursing notes and discharge summaries are recorded in free-text fields in the database. Nurses can enter any information here. ECG, echo and radiology reports are also available.
- Chart: A patients medical chart contain any parameters recorded by the staff: validated physiologic recordings, demographic information, weight, height and ventilator settings are examples of the type of information recorded here.
- Laboratory tests: Results of blood gas, chemistry and other body fluid tests are recorded here.

Demographics

When patients arrive at the hospital, their demographic information is recorded. This information is stored in the d_demographicitems and demographicevents tables. The view demographic_detail is provided for convenience.

Items

The items tables record the items which can be recorded for a particular event. As such, they are related to corresponding events and durations tables. Each item has a unique itemid, as well as a label and a category.

- D_MedItems
- D_ChartItems
- D_IOItems
- D_ParamMap_Items
- D_DemographicItems
- D_Codeditems

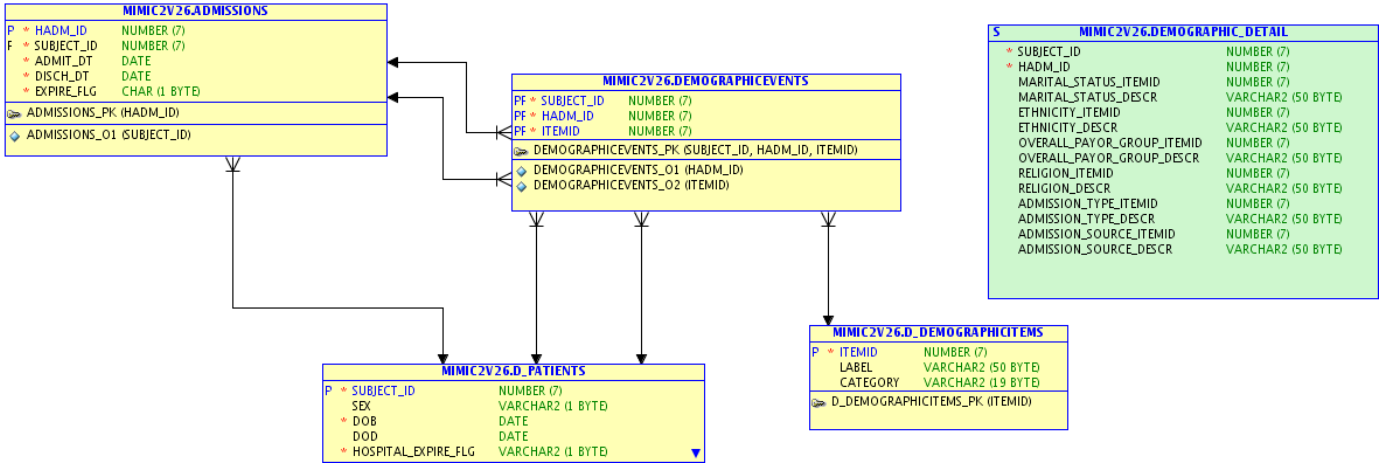


Figure 2.8: The demographic information recorded for a patient's hospital admission.

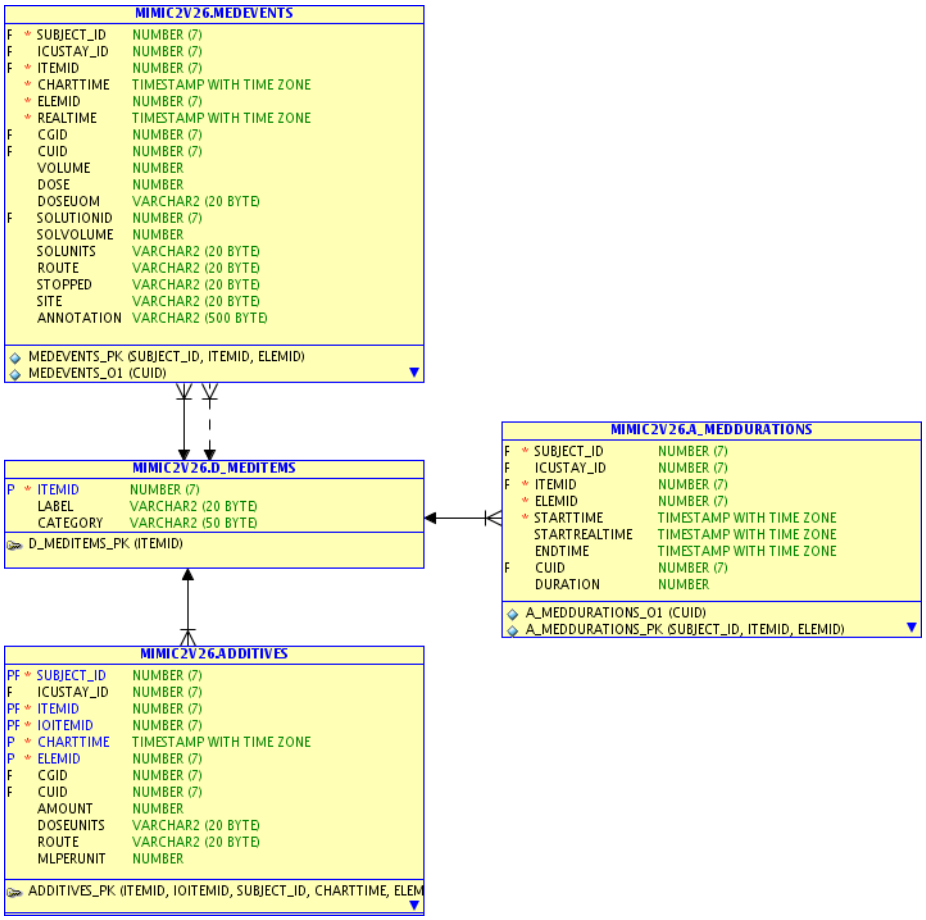


Figure 2.9: Patient medication is stored in 4 tables. The medevents, d_meditems, a_meddurations and additives tables record all data related to patient medication.

Medications

Medication(s) given to a patient are recorded in the medevents, d_meditems, a_meddurations and additives tables. The relationships are shown in Figure 2.9.

The tables contain details of the available drugs and information related to the particular administration. The med_events table contains information pertaining to the patient, dosage and annotation. The additives table contains information related to the additives which are included with the drug administration. The a_meddurations table contains information about the time and duration of the medication and the d_meditem table links the other tables together.

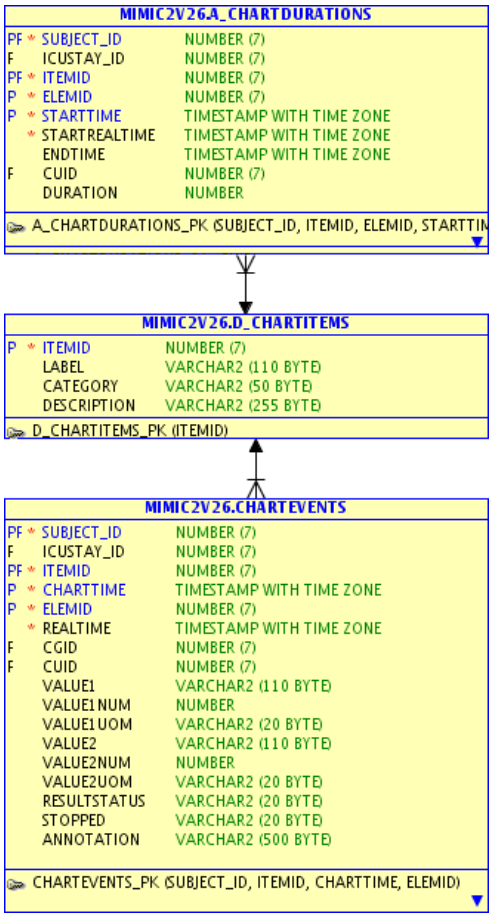


Figure 2.10: Patient chart data is stored in 4 tables. The chartevents, d_chartitems, a_chartdurations and form events tables record all data related to patient charts.

Charts

Patient medical chart data is recorded in the chartevents, d_chartitems, a_chartdurations and formevents tables. The relationships are shown in Figure 2.10.

Fluids

Patient input/output (IO) data is recorded in the ioevents, d_ioitems, a_iodurations, deliveries, totalbalevents and additives tables. The relationships are shown in Figure 2.11.

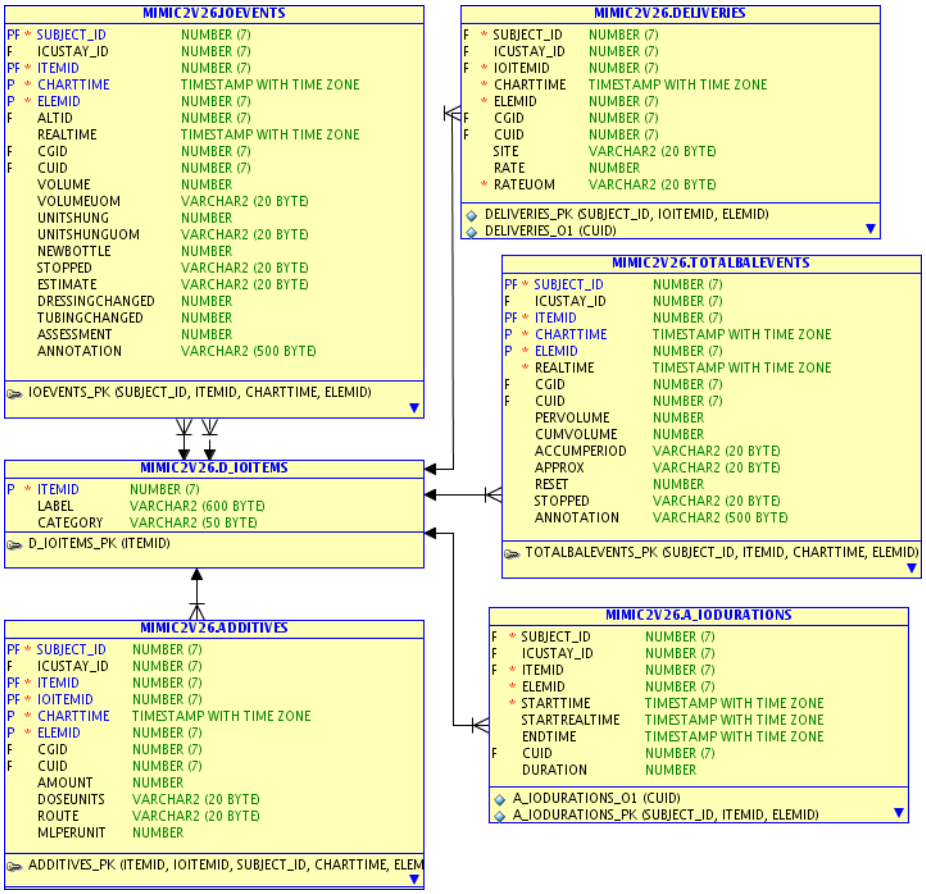


Figure 2.11: Patient IO data is stored in 6 tables. The ioevents, d_ioitems, a_iodurations, deliveries, totalbalevents and additives tables record all data related to patient charts.

Notes

Patient notes are recorded in the noteevents table. The relationships are shown in Figure 2.12.

During the course of the patient stay in the ICU, free text "notes" are produced by the hospital staff. Nurses typically write "nursing notes", a summary of events which occurred during their shift period. When the patient is discharged from the hospital, the responsible physician dictates a summary of the entire hospitalization period, known as the "discharge summary". These reports are recorded in the MIMIC II database.

Progress or Nursing notes and discharge summaries are stored in the noteevents table and are linked to patients through the *Subject_ID*. Notes are

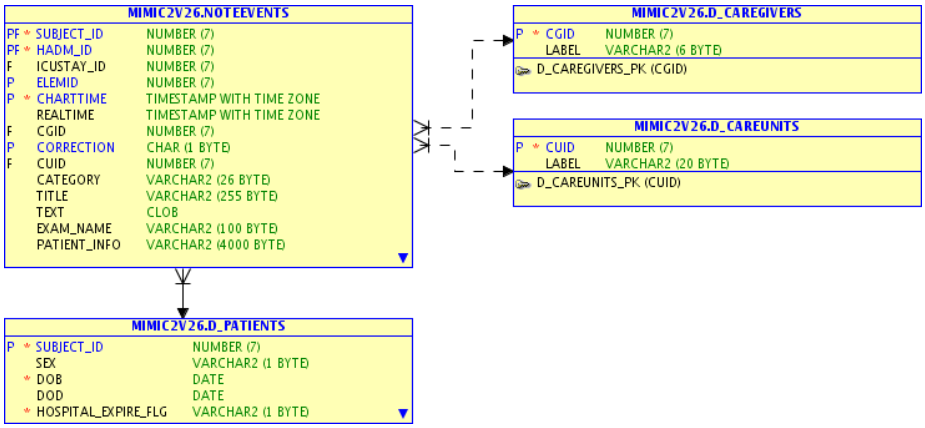


Figure 2.12: Patient notes/reports are stored in the noteevents table.

linked to care_givers and care_units through their unique identifiers. The noteevent table also contains relevant meta-data such as the data and time of entry and various other categories and IDs. A sample discharge summary has been shown previously, in Section 1.7. The data is free text and can contain anything entered by the nurse.

The noteevents table also contains reports from various diagnostic tests. Reports from X-rays, echos and ECGs are found in the noteevents table. The table also contains relevant meta-data such as the data and time of entry and various other categories and IDs.

Figure 2.13 shows a sample radiology report. The data is free text and contains information obtained from radiology.

Procedures

Procedures performed on a patient are recorded in the procedureevents table. Its relationships with the other tables are shown in Figure 2.14.

Laboratory and microbiology tests

Laboratory and microbiology tests are performed throughout a patients hospital stay. The database tables containing the results of these tests are shown in Figure 2.15.

Waveform metadata

Metadata from the high resolution waveforms and trends (Section 2.3) is also included in the relational database. The data found in these tables (Figure 2.16) is roughly analogous to the output obtained by running "wfdbdesc" (Section 2.3.2) from the WFDB toolbox on all of the available waveforms.

```
Reason: CHECK ETT TUBE PLACEMENT, ?PNA, CHF
[**Signature 1**]
UNDERLYING MEDICAL CONDITION:
 85 y/o male s/p acute mi and catherization now in ccu with
cardiogenic shock
REASON FOR THIS EXAMINATION:
CHECK ETT TUBE PLACEMENT
?PNA
CHF
[**Signature 1**]
 FINAL REPORT
CLINICAL INDICATION:  Assess endotracheal tube placement in
patient with congestive heart failure.

Comparison is made to previous study of one day earlier.
An endotracheal tube is present, in satisfactory position.
A Swan-Ganz catheter terminates in the proximal left
pulmonary artery and has been withdrawn in the interval.
An intraaortic balloon pump terminates about 3.3 cm below
the superior aspect of the aortic knob, and a nasogastric
tube terminates in the region of the gastroduodenal
junction.

Cardiac and mediastinal contours are stable in the interval
and pulmonary vascularity is within normal limits for
technique.  There has been improvement in the left
retrocardiac opacity and there remains a patchy right
basilar opacification which is slightly increased.  A small
amount of fluid is seen in the minor fissure.

IMPRESSION:

1) Lines and tubes in satisfactory position, as detailed
above, with no evidence of pneumothorax.

2) Improved left retrocardiac opacity and worsened right
lower lobe opacity likely due to atelectasis.

JPE

   DR. [**First Name11 25**] [**Initials 5**] [**Last Name
26**]Approved: SAT [**13-09-01**] 7:27 PM
```

Figure 2.13: Sample radiology report. The text field of the radiology report table contains information obtained from radiology.

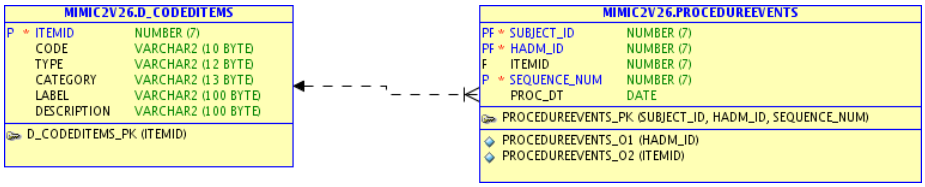


Figure 2.14: Procedures performed on a patient. The d_codeditems and procedureevents tables record all data related to patient procedures.

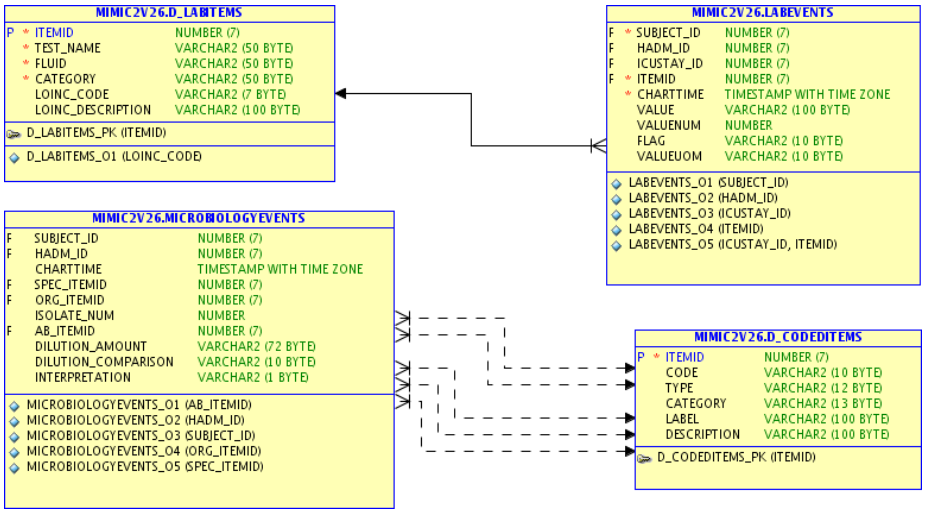


Figure 2.15: Laboratory and microbiology tests. These data are stored in the labevents and microbiologyevents tables. d_labitems and d_coded items contain full descriptions of the lab tests (with LOINC codes) and microbiology tests (specimin, organism and antibiotic).

2.2.6 Summary

All of the information described so far can be thought of "discrete" patient data. It is generally recorded manually and only requires infrequent updates. For example, admission/discharge only occurs once during a patient stay. ICU transfer will only occur a few times. Medication will only occur a few times a day and reports will be added when particular diagnostics are performed.

In contrast, the high resolution waveforms which are discussed next are recorded constantly. Measurements are recorded automatically by computer 125 times per second. The relational database described above is a poor device for recording data of this type and a separate system is used to store these waveforms.

Figure 2.16: Waveform Metadata. These table contain metadata for the high resolution waveforms available on PhysioNet. The Subject_ID column contains the related subject for the particular waveform. These tables simplify searching for subjects in the relational database who also have corresponding high resolution waveform records

2.3 High resolution waveforms and associated trends

2.3.1 Overview

High resolution waveforms were converted from a proprietary format into MIT's WFDB format (explained later in this chapter). Our waveform collection efforts were essentially broken into two groups, which reflect different versions of the Philips waveform archiving software. The first group includes approximately 2,800 records, each consisting of:

- Up to 4 simultaneous channels sampled at 125 Hz, with an amplitude resolution of 8 bits, normally containing 2 leads of ECG, ABP and a pulmonary artery pressure (PAP) if available. If the PAP was not available, a respiratory signal (impedance pneumogram) an oxygen saturation waveform (photoplethysmogram) was recorded.
- Up to 30 parameters recorded once a minute with an amplitude resolution of up to 16 bits . Parameters include heart rate (HR), systolic/mean/di-

astolic blood pressure (ABPSys/ABPMean/ABPDias), peripheral oxygen saturation (SpO2), and cardiac output (CO).

- All bedside monitor-generated alarms (such as arrhythmias) and in-ops (such as sensor disconnects).

Upgrades in hardware and software allowed us to improve data collection and gather a second group of records to include:

- Up to 8 simultaneous channels sampled at 125 Hz.
- Analog to digital resolution of 10 bits.
- Trend parameters with a temporal resolution of once a second.

This second, improved, data collection effort is ongoing, and will be added to the public database as we process them.

2.3.2 The WFDB software package

Over the past twenty years, the team at PhysioNet has developed a large collection of open source software to store, analyze and manipulate physiological measurements Goldberger et al. [2000].The WFDB software package is written in highly portable C and can be used on all popular platforms, including GNU/Linux, Mac OSX, MS-Windows, and all versions of Unix. A set of wrappers allow the integration of the WFDB library with other programming languages and interfaces so that the tools can also be run from within visualization tools or other programming environments. Table 2.3 summarizes the major components of the WFDB Software Package; a more detailed description is available at the PhysioNet web site (`http://www.physionet.org/`).

2.3.3 MIMIC II waveform records

All MIMIC II waveforms are stored in WFDB format. Table 2.3 summarizes the types of files you will find for the MIMIC II waveform database.

The records vary in length; some are several weeks in duration. It is common for the signal sources to be interrupted or changed occasionally during recordings of such a long duration. In a typical waveform database, you will find a directory layout including several record names or "cases". All files associated with each record are gathered in a sub-directory named after the record. For example, the files associated with record a40001 are all located within the directory named a40001.

Table 2.4 presents some useful WFDB commands to navigate through the waveform records. In a typical WFDB database record, a header file specifies the names of the associated signal files and their attributes, briefly:

- Record name: a string of characters that identifies the record.

Component	Description
WFDB library	This is a set of functions for reading and writing files in the formats used by PhysioBank databases. The library supports reading directly from remote servers allowing applications linked with the WFDB library to view or analyze data without the need to download entire records and to store them locally.The WFDB library is implemented in C but provides interfaces for software written using Perl, Python, C# (and other .NET languages), Java, Matlab, PHP, Ruby, TCL, and several versions of Lisp.
WFDB applications	A large set of well-tested, interoperable command-line tools for signal processing and automated analysis.These applications are described in detail in the WFDB Applications Guide Goldberger et al. [2000].
Visualization tools: WAVE	Extensible interactive graphical environment for manipulating sets of digitized signals with optional annotations

Table 2.3: WFDB software package major components

File type	Extension	Description
Header	.hea	Contain signal file names and attributes in plain text format.
Signal	.dat	Contain signals binary data.
Annotations	-dependent, e.g. ".al" for alarms, ".wqrs" for ECG beat annotations	Contain signal custom annotations in binary format
Calibration	.cal	Contain signal calibration specifications.

Table 2.4: WFDB file types

Command	Description
wfdbdesc	Reads specifications for the signals described in the header file for record.
rdsamp	Reads signal files for the specified record and writes the samples as decimal numbers on the standard output. Each line of output contains the sample number and samples from each signal, beginning with channel 0, separated by tabs.
wave	Can be used to view the specified WFDB record or records on any display controlled by an X11 server. It includes facilities for interactive annotation editing. The keyboard and mouse are used to control the display interactively

Table 2.5: Useful WFDB commands

- Number of channels: an integer greater than zero.
- Sampling rate: an integer or floating-point number, interpreted as samples per second.

Figure 2.17, shows the output of the "wfdbdesc" command, which outputs a human-readable description of a waveform record.

If you want to display the signal contents of a particular record, you can use the "rdsamp" command. Figure 2.18 shows the output of this program.

Another option is to use a visualization tool like WAVE, or ATM (`http://www.physionet.org/cgi-bin/ATM`), to display the contents of a particular waveform record. Figure 2.19 shows the display of a particular waveform record. Note that the signal processing algorithms can be run from this viewer.

2.3.4 Alarms and Inops

The following description about alarms and inops corresponds to the first group of waveforms collected (approximately 2,800 patient records). All are adults where the *Case_ID*'s are numbered less than a44000.

Simultaneously with approximately 10,000 patient-days of waveforms and trends, we have collected over 450,000 alarms and inops. This amounts to a frequency of one alarm or alert every thirty minutes for each patient in the hospital, although most of these are not life-threatening. Tables 2.6 and 2.7 list the types of alarms and *inops* (non-physiological alerts such as machine disconnections) gathered by our data collection system. A three-star alarm is a potentially life-threatening condition that requires immediate attention. Two-star alarms require less immediate attention, although may provide warning of an increased risk of adverse problems in a patient over time. Note that most of these alarms are not verified as correct. In particular, the *.alarm* are all unverified. A subset of alarms (with the *.alM* extension) are verified by humans, and are described in the next section.

```
Starting time: [16:24:28.848 30/03/2011]
Length: 1:28:00.000 (660000 sample intervals)
Sampling frequency: 125 Hz
4 signals
Group 0, Signal 0:
  File: a42174\_000006.dat
  Description: II
  Gain: 55 adu/mV
  Initial value: 1
  Storage format: 80
  I/O: can be unbuffered
  ADC resolution: 8 bits
  ADC zero: 0
  Baseline: 0
  Checksum: 19538
Group 0, Signal 1:
  File: a42174\_000006.dat
  Description: V
  Gain: 39 adu/mV
  Initial value: 4
  Storage format: 80
  I/O: can be unbuffered
  ADC resolution: 8 bits
  ADC zero: 0
  Baseline: 0
  Checksum: -9315
Group 0, Signal 2:
  File: a42174\_000006.dat
  Description: ABP
  Gain: 1.25 adu/mmHg
  Initial value: 184
  Storage format: 80
  I/O: can be unbuffered
  ADC resolution: 8 bits
  ADC zero: 0
  Baseline: -100
  Checksum: -3865
Group 0, Signal 3:
  File: a42174\_000006.dat
  Description: PAP
  Gain: 2.5 adu/mmHg
  Initial value: 125
  Storage format: 80
  I/O: can be unbuffered
  ADC resolution: 8 bits
  ADC zero: 0
  Baseline: -100
  Checksum: -4501
```

Figure 2.17: Sample output for wfdbdesc

Annotated alarms

Since no large annotated dataset of alarms is publicly available, a set of gold standard alarms to support the development and testing of a false alarm suppression algorithm was generated from the above alarms. Initially we have concentrated on life-threatening arrhythmia alarms, namely; Asystole, Extreme Bradycardia, Extreme Tachycardia, Ventricular Tachycardia and Ventricular Fibrillation. In order to assemble such a database we first searched for patient records in the MIMIC II database that met the following two criteria:

1. Record contains at least one of the 5 above listed critical alarm categories.
2. At least one of the alarms is associated with simultaneous ABP and ECG waveforms.

Our initial search yielded 496 patient records with a total of 45,370 hours of simultaneous ECG & ABP waveforms containing 8,636 alarms. Each alarm was manually reviewed by two independent experts, and discrepancies were adjudicated by a third expert.

```
time     II     V      ABP     PAP
(sec)   (mV)   (mV)   (mmHg)  (mmHg)
-----------------------------------------
0.000   0.018 0.103 22.400 90.000
0.008   0.018 0.077 22.400 90.000
0.016   0.036 0.051 22.400 90.000
0.024   0.018 0.051 22.400 90.000
0.032 -0.018 0.051 60.800 90.000
0.040   0.000 0.026 60.800 90.000
0.048 -0.018 0.026 60.800 90.000
0.056 -0.036 0.000 60.800 90.000
0.064 -0.036 0.026 22.400 90.000
0.072 -0.018 0.000 22.400 90.000
0.080 -0.055 0.000 22.400 90.000
0.088 -0.073 0.000 22.400 90.000
0.096 -0.055 0.000 60.800 90.000
0.104 -0.055 0.000 60.800 90.000
0.112 -0.036 0.000 60.800 90.000
0.120 -0.036 0.000 60.800 90.000
0.128 -0.055 0.000 22.400 90.000
0.136 -0.073 0.000 22.400 90.000
0.144 -0.073 0.000 22.400 90.000
0.152 -0.073 0.000 22.400 90.000
0.160 -0.073 0.000 60.800 90.000
0.168 -0.073 0.000 60.800 90.000
0.176 -0.073 0.000 60.800 90.000
```

Figure 2.18: Sample output of rdsamp

Alarm repetitions referring to the same event, were removed. Furthermore, all 48 patients that possessed active intra-aortic balloon pumps (IABP) were removed, since their ABP waveforms do not appear as "physiologically normal". The final set comprises 448 patients with 5,386 alarms with simultaneous ABP & ECG waveforms. These annotations have been posted on PhysioNet with the file extension *.alM*.

Full details of how these alarms were annotated is available in Aboukhalil *et al* Aboukhalil et al. [2008], together with an evaluation of their statistics.

2.3.5 Signal Quality

Although our search engines allow a researcher to determine what signals exist for which patients, this is no guarantee of quality, and sometimes the data can be so noisy that no useful clinical information can be extracted from the data. To avoid requesting noisy data, and using this data for further processing, we have developed a set of signal quality indices (SQI's) for both electrocardiogram and blood pressure data. We have also used these indices and a multi-channel

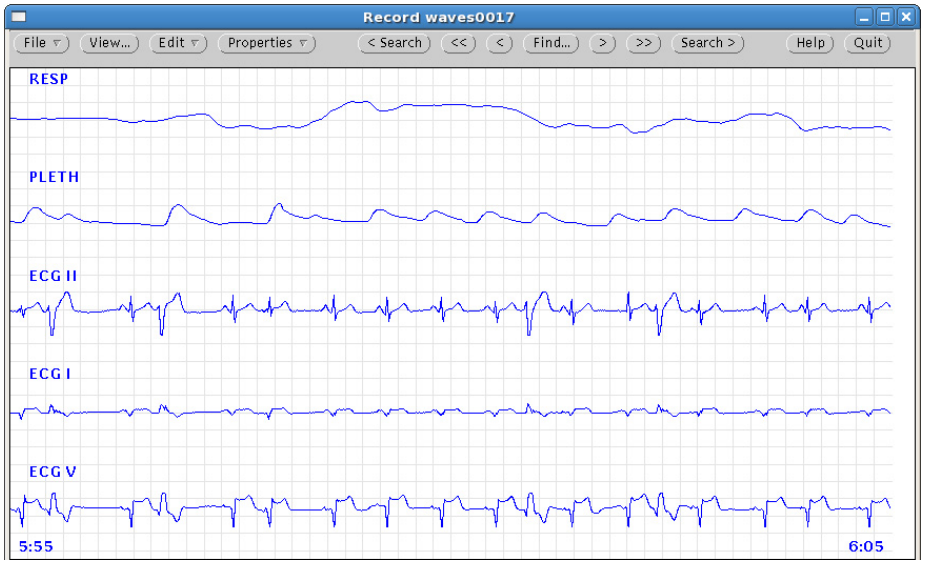


Figure 2.19: Sample waveform record

weighting algorithm, to generate our best estimates of heart rate and blood pressure for every 10 second window of waveform data.

The ECG signal quality metrics are a combination of statistical measures, in both the time and frequency domains, multi-channel QRS detector performance, and correlation to past data. A more in-depth description of the formation of these annotations, together with an evaluation of a robust HR and ABP tracking algorithm that utilizes this information can be found in Li et al. [2008] and Li et al. [2009]. The blood pressure signal quality metric is based upon two earlier developed metrics by Zong Zong et al. [2004] and Sun Sun et al. [2006].

There are several annotations associated with each beat in the ECG and ABP signals. Table 2.8 describes the available SQI annotations for waveform records. Full descriptions of how to interpret the SQI output for ECG and ABP can be found in in Li et al. [2008] and Li et al. [2009] respectively. Note that although many of the artifact types for each of these signals have been incorporated, and known errors in heart rate and blood pressure calibrated to the SQI output, some artifacts are not well represented. In particular, the tricky problem of blood pressure damping is not yet fully solved in our ABP SQI metric. Any analysis of blood pressure should therefore be tempered by the fact that damping may lead to an error, and in particular, an under-estimation of the SBP and pulse pressure.

Table 2.8: Per beat SQI annotations files. Where "file" is the waveform record name such as "a41000". SDR = Spectral Distribution Ratio. DF = digital filter. LT = length transform. EPLTD and WQRS are the beat detectors that

Alarm Label	Definition
***ASYSTOLE	No QRS for 4s
***BRADY m < n	Bradycardia < 40BPM
** HR m < n	Low Heart Rate
** HR m > n	High Heart Rate
** IRREGULAR HR	Irregular Heart Rate
** MISSED BEATS	Missed Heart Beats
** MULTIFORM VPBs	Multiform Ventricular Premature Beats
** PACER NOT CAPT	Pacemaker not capturing
** PAIR VPBs	Pair of Ventricular Premature Beats
** R-ON-T VPBs	R on T type beats
** RUN VPBs 3 - 9	Run of 3-9 Ventricular Premature Beats
** RUN VPBs > 9	As above, but >9
** STi m.m<n.n	ST depression in mV
** STi m.m>n.n	ST elevation of in mV
***TACHY m > n	Tachycardia, HR > n BPM
** VENT BIGEMINY	Ventricular Bigeminy
***VENT FIB/TACH	Fibrillatory Waveform for 4s or more
** VENT RHYTHM	Ventricular Rhythm
***VENT TACHY	Run of >=5 Ventricular Beats with HR>100
** VENT TRIGEMINY	Ventricular Trigeminy

Table 2.6: ECG alarms in the MIMIC II database together with their definitions.

use the DF and LT methods respectively. See Li et al. [2008] and Li et al. [2009] details.

Annotation	**Description**
file.ecgsqid	A combined beat-by-beat ECG SQI created by selecting the best ECG SQI between different ECG leads.
file.ecgsqid$n$	The ECGSQI annotation of the n^{th} $(n = 0, 1, 2, ...)$ ECG lead, used to create a combined annotation file'.multid'

continued on next page

Alarm Label	Definition
***ABP m < n	Hypotension (Extremely Low Blood Pressure)
** ABP m < n	Low Blood Pressure
** ABP m > n	High Blood Pressure
***APNEA	No Respiratory Effort Detected
** ART m < n	Low Blood Pressure (secondary line)
** ART m > n	High Blood Pressure (secondary line)
** CVP m < n	Low Central Venous Pressure
***Desat m < n	Desaturation (of SPO2) (#)
** ICP m > n	Low Intra-Cranial Pressure
** LAP m > n	High Left Arterial Pressure
** NBP m < n	Low Non-Invasive Blood Pressure
** NBP m > n	High Non-Invasive Blood Pressure
** PAP m < n	Low Pulmonary Arterial Pressure
** PAP m > n	High Pulmonary Arterial Pressure
** P1 m < n	Low (Generic) Pressure
** RESP m > n	Hyperventialaation
**SpO2 m < n	Low Oxygen Saturation
**Tblood m.m<n.n	Low Blood Temperature (in deg C)
** UAP m < n	Low Umbilical Arterial Pressure
** UAP m > n	Low Umbilical Arterial Pressure
** UVP m > n	High Umbilical Venous Pressure
***ABP DISCONNECT	Invasive Arterial Line Disconnect
***PAP DISCONNECT	Pulmonary Arterial Catheter Disconnect
***UAP DISCONNECT	Uterine Pressure Line Disconnect

Table 2.7: Non-ECG alarms and Inops in the MIMIC II database. (#) indicates that this alarm is available only on the latest version (revision F) of the bedside monitor software.

continued from previous page

Annotation	**Description**
file.epltd$n$	The EPLTD (DF) annotation of the n^{th} ($n = 0, 1, 2, ...$) ECG lead, used to create EPSQI and ICHSQI
file.epsqi$n$	The EPSQI (bSQI) annotation of the n^{th} ($n = 0, 1, 2, ...$) ECG lead, used to create Kurtosis (kSQI) and SDR (sSQI)
file.ichsqi$n$	The ICHSQI (iSQI) annotation of the n^{th} ($n = 0, 1, 2, ...$) ECG lead, used to create ECGSQI

continued on next page

continued from previous page	
Annotation	**Description**
file.kurtd$n$	The Kurtosis (kSQI) and SDR (sSQI) annotation of the n^{th} ($n = 0, 1, 2, ...$) ECG lead, used to create ECGSQI and the sample-and-hold HR (HRsh1) of the first ECG lead
file.multid	An all-in-one annotation include all leads of ECG beats with ECGSQI and ABP beat with ABPSQI, used to create the HR and ABP
fileTd	A WFDB trend data file that includes different calculations of HR, HRSQI and ABP sampled at 0.1Hz
fileTd.hea	A header file of fileTd
file.wqrs$n$	The WQRS (LT) annotation of the n^{th} ($n = 0, 1, 2, ...$) ECG lead, used to create EPSQI (bSQI)
fileT	A WFDB trend data file similar to fileTd, but without baseline wander filter to calculate kurtosis.
fileT.hea	Header file for fileT
file.wsqi	The WSQI annotation of the ABP lead, used to create ABPSQI
file.jsqi	The JSQI annotation of the ABP lead, used to create ABPSQI
file.abpsqi	The ABPSQI annotation of the ABP lead, used to create '.multid'

Chapter 3

Database Access

3.1 Introduction

This chapter provides instructions for connecting to and extracting data from the MIMIC II database. There are two methods for accessing the data:

- Flat file download from PhysioNet http://physionet.org
- Web-based "MIMIC Explorer" http://mimic.physionet.org/

In order to gain access to the database, you must follow the instructions posted on PhysioNet (http://www.physionet.org/physiobank/database/mimic2cdb/restricted/)

When you have completed and signed the data use agreement, your application will be processed and your access details emailed to your account.

Once you have completed this process, an account will be created for you to access the restricted areas of the MIMIC II website and the MIMIC Explorer.

Chapter 4

Examples of data analysis

4.1 Introduction

This chapter provides some introductory examples for obtaining data/statistics from the database. We hope that these examples will enable users to become familiar with the tables and the data they contain. These examples will also help to illustrate the complexity of the database and the difficulties encountered in obtaining certain data. Please note that the examples shown here have been tested on database version MIMIC2V25. Although direct copy-and-paste of the examples is possible, some examples may result in "Invalid Character" errors. For example, the asterisk (*) character may not be copied-and-pasted correctly. In this case, simply type the correct character in its place.

4.2 Clinical Examples

4.2.1 Patient population age statistics

The first example is simple. Query 4.1 simply counts the number of unique subject ids in the database.

Listing 4.1: SQL to obtain the number of subject ids in the database

```
-- Subject IDs
  select count(*)
    from d_patients
```

We have developed a database table which contains a large number of columns and provides lots of summary data. This table is called "*icustay_detail*" and contains information relating to patient stays in the ICU, their hospital ad-

missions and various other paramters. This data can be obtained by running query 4.2

Listing 4.2: ICU stay detail table

```
-- ICU Stay Details
  select *
    from icustay_detail
```

The result of the above query contains many details about patients and their ICU stays. There are columns which provide the number of admissions and ICU stays for each patient, DOB, admission and discharge dates, flags indicating whether or not the patient died, gender, and finally, basic statistics including weight, height and SAPS score.

4.2.2 Resolving discrepancies between multiple itemIDs for one parameter

Since the database contains many "free text" itemIDs, there is no unique method for representing certain parameters. For example, CPR is mentioned in 9 different itemIDs. Each itemID can contain a variety of values such as "yes", "no", "done", "performed", etc. In addition multiple different capitalization and spelling errors are found. This makes it extremely difficult to obtain accurate information on whether or not CPR was performed for a particular patient.

As and when certain data is extracted from these "free text" fields, it can be moved into more meaningful fields which will permit simpler data extraction. For example, the CPR itemIDs mentioned above could be translated into a binary field which simply states whether or not CPR was performed.

Examples of multiple mappings, plus code to merge them, can be found in appendix 6.2.

Chapter 5

Quick-Start, Frequently Asked Questions and Known problems/issues

5.1 Quick Start

There really is no quick start to exploring the data in MIMIC II DB. First you need to decide which type of data you want to process:

1. 'waveform' data
2. high (temporal) resolution trend data (numerics)
3. 'clinical' data - low temporal resolution data and/or categorical data.

You may also want to decide if you want to use all of the available patients, or just narrow your focus on a subset cohort. For example, you may wish to concentrate on patients who were admitted with a particular diagnosis, or who were administered with a particular drug. In order to find patient cohorts, you may wish to use our prototype search engine, which can be found at:
http://mimic.physionet.org.
Currently, with this interface you can search in the following ways:

- Locate free text strings the nursing progress notes and discharge summaries.
- Locate patients with particular combinations of waveforms.
- Locate patients with particular demographics (age and gender).

Each of these queries can be 'ANDed' together to construct more restrictive subsets.

Once you have found your subset of patients on which you want to work, you will want to download the data you wish to work on, or perform more detailed queries. Methods for doing this are described in chapter 3. Sample queries can be found in chapter 4. We strongly suggest you work through these before attempting to construct patient cohorts and analyze data. We also strongly advise that you take note of the different definitions of patient identifiers, as described in section 1.4.2. Note that you can analyze patients by individual (subject) ID, hospital admission, ICU stay, or by waveform Case ID. We also suggest that you read the frequently asked questions (FAQs) below.

5.2 FAQs about the MIMIC II database

1. **How can I get an answer to my question?** Please ensure that your question has not already been answered by this document; either in the main body of text or in these FAQs. If you are sure that your question is not addressed by this document, then please email the authors with your question.

2. **What is the MIMIC II database?** The Multiparameter Intelligent Monitoring in Intensive Care II database Moody and Mark [1996] contains detailed clinical data from patients hospitalized in Intensive Care Units (ICUs).

3. **Where is the MIMIC II database physically located?** The MIMIC II database is hosted by MIT: it is located on the Cambridge Campus.

4. **Who do I contact for more information?** Current contact information is always available on the MIMIC II website (`http://mimic.physionet.org/`)

5.3 FAQs about data access

1. **What methods can I use to access the MIMIC II database?** There are 3 access methods for the MIMIC II Database: Oracle's SQL Developer, JDBC bib and WFDB. In addition, complete database dumps are available for import into your local systems; such data can then be accessed in any desired method. Please see Chapter 3.1 for more details on data access methods.

2. **How do I gain access to the MIMIC II database?** You will need to agree to and sign our data use agreement in order to gain access to the database. You will also have to fill out a short form detailing your plans for the data. Simply visit our website (`http://mimic.physionet.org`) to apply for access and to obtain further information.

3. **Who can access the MIMIC II database?** Anyone who wishes to perform research on the MIMIC II data will be permitted access to the database. Simply visit our website to apply.

5.4 Known issues/problems

1. **Waveform-trend misalignment.** Although the trends should match up with the parameters derived from the waveforms, this is not always true. This can be due to filter delays, network timing errors or data server timing errors.

2. **Inter-waveform alignment problems.** The method used for MIMIC waveform data extraction was not designed for inter-waveform analysis. The waveform data contain unspecified/unknown filtering delays and/or unknown inter-channel delays, which may not be constant in a given record. Therefore, although the ECGs are time-aligned, there may be a (changing) delay of up to 500ms between any of the other waveforms in the data. Therefore, no pulse transit times can be accepted to be true (absolute or relative).

3. **Missing waveform and trend data.** Every patient will have some level of data missing between the admit and discharge time. This can be for many reasons:

   - The patient was disconnected from the monitor for some period (perhaps for a scan or to replace electrodes).
   - The data collection system or the network over which the data was transmitted crashed.
   - The data that was collected was corrupted and conversion to WFDB was not possible.

4. **The clinical data change dimensionality over time and between patients, and are irregularly and sparsely sampled.** The amount and type of data that are recorded concerning a patient, and the frequency at which it is sampled is a function of both the settings on the monitoring equipment, and the activities of the clinical staff. This in turn is reflective of the clinical team's understanding of the patient's changing condition(s). Many tests are not routine and therefore are only ordered when the clinical team suspects a given condition based on the presenting observations. Therefore, the dimensionality of the data for a given patient may fluctuate over time and no signal is guaranteed at any given point in time. When a patient's condition becomes more acute, data are often sampled more frequently, and the number of sampled parameters increases. This leads to the question of what to do with missing data. Interpolation and imputation schemes generally perform poorly because there are no models of *how* the data are missingAbdala and Saeed [2004]. It should also be

noted that prediction or classification algorithms can be 'fooled' by the *presence* or *absence* of a data stream. That is, it may not be the result of a test that causes an algorithm to give a particular output, but rather just the fact that a clinician thought the particular test was needed. Caution should be taken in the interpretation of such results.

5. **Contradictory data.** Some data derived from the waveforms or trends may be incommensurate with each other, or with the data in the relational database. This can be due to noise in the data, the use of different windows and filters to process the data, time alignment errors, or the fact that humans can override the machine transmitted data (in the relational database) with values that they think more correctly reflect the patient's physiology. It should be noted, that these cannot always be trusted Hug and Clifford [2007].

6. **Multiple data streams / itemIDs for a single parameter.** Each parameter may be recorded in a variety of ways by both humans and machines. For example, the heart rate (HR) can be derived from the ECG, ABP and PPG (pulse oximeter). You should not expect these parameters to give the same exact values. They will also respond to artifacts in different manners, and sometimes be affected at different points in time by the artifacts.

   In the relational database, each signal or parameter may be recorded under a variety of different names. For example, Lactic acid values are found in chartevents-818 and chartevents-1531. A current list of the known mappings can be found in Appendix 6.2, although we encourage users to send us other mappings that they discover.

7. **Possible mistakes in the subject-case ID mapping.** Linking data from the bedside monitors and the other hospital databases was not a trivial process. Although names and medical record numbers are sometimes manually entered into the bedside monitor, often they are not, or are done so incorrectly. Furthermore, even when a patient is discharged from the ICU, they are sometimes not 'discharged' from the bedside monitor, and so the next patient may inadvertently inherit the name and MRN of the old patient. Therefore, one should be attentive to this possibility. For the patients with no MRN or name identifiers in the waveform and trend data, we attempted to match the patients based on admit/discharge times, available trends, and numerics of the data. This form of matching is obviously more error prone than MRN or name matching. See section 1.4.3 for more information.

   Although every effort has been made to map the waveform and trend data to the associated clinical data, mistakes will be present. If you think you have discovered such a mistake, please email us with the evidence and we will do our best to answer your query or correct the data.

8. **Possible mistakes in calibration or conversion units** Care should be taken to identify data that appears to be out of range or exhibiting abnormal offsets. For example, temperature may be measured in degrees Centigrade and recorded in degrees Fahrenheit for part of a patient's record. More fundamentally, conversion factors may have become corrupted, and so representations of parameters may not always be correct.

9. **Possible mistakes in waveform labels** We have noticed that in converting to an open format, the data, which was written to disk in a proprietary format using Microsoft .Net, errors have crept into the waveform labeling. Sometimes channels labelled as V (ECG) are actually respiratory waveforms. At other times, labels are "UNKNOWN" and although they are often PPGs, this is not always true.

10. **My drug is having the opposite effect of what I expected** Drugs effects are variable, depending on interactions with other drugs, dosage levels and cardiovascular state. See section 6.2.17

11. **The nursing note does not make complete sense or contradicts the data**. Nursing notes are 'free-text' notes that can contain typos, errors or hard to understand short-hand. While we have tried to provide a list of useful abbreviations in section 5.5, this is not complete and errors may still exist. Note also that the numerical data may be in error.

We are always striving to improve our database, and so if you notice any anomalies, and/or have any suggestions on how to fix them, please do contact us.

5.5 Abbreviations

- **A** - Assessment
- **ABP** - Arterial Blood Pressure
- **ABPSQI** - Arterial Blood Pressure Signal Quality Index
- **BIDMC** - Beth Israel Deaconess Medical Center
- **BP** - Blood Pressure
- **BPM** - Beats (or Breaths) per Minute
- **BUN** - Blood Urea Nitrogen (also known as Urea or Urea nitrogen)
- **CAREVUE** - The Philips bedside ICU workstation for clinicians
- **CCU** - Cardiac Care Unit
- **CDSS** - Clinical Decision Support System
- **CMO** - Comfort Measures Only

- **CO** - Cardiac Output
- **CSRU** - Cardiac Surgery Recovery Unit
- **CVP** - Central Venous Pressure
- **DBP** - Diastolic Blood Pressure
- **DF** - Digital Filter
- **DNI** - Do Not Intubate
- **DNR** - Do Not Recusistate
- **D/C'd** - Discontinued, or Discharged
- **ECG** - Electrocardiogram
- **ECGSQI** - Electrocardiogram Signal Quality Index
- **ECO** - Estimated Cardiac Output
- **EKG** - Electrocardiogram
- **F/E** - Fluid and Electrolytes
- **GCS** - Glasgow Coma Scale (or sometimes *Score*
- **GI** - Gastrointestinal
- **HEME** - Hematology
- **HIPAA** - Health Insurance Portability and Accountability Act
- **HR** - Heart Rate
- **IBP** - Invasive Blood Pressure
- **IABP** - Intra-Aortic Balloon Pump or Invasive Arterial Pressure
- **ICD** - Implantable Cardioverter Defibrillator
- **ICD-9** - International Statistical Classification of Diseases and Related Health Problems (version 9)
- **ICU** - Intensive Care Unit
- **ID** - Identifier or Infectious Disease
- **ISM** - Information Support Mart
- **LT** - Length Transform
- **MBP** - Mean Blood Pressure

- **MICU** - Medical Intensive Care Unit
- **MRN** - Medical Record Number
- **NBP** - Non-invasive Blood Pressure
- **Neo** - Neosynephrine
- **NIBP** - Non-invasive Blood Pressure
- **NICU** - Neonatal Intensive Care Unit
- **NSR** - Normal Sinus Rhythm
- **P** - Plan
- **PAP** - Pulmonary Arterial Pressure
- **PCWP** - Pulmonary Capillary Wedge Pressure (or *Wedge Pressure*)
- **PPG** - Photoplethysmogram
- **PRBC** - Packed Red Blood Cells
- **RR** - Respiration Rate
- **SAPS** - Simplified Acuity Score
- **S/0** - Sign Out
- **SaO2** - Arterial Oxygen Saturation
- **SBP** - Systolic Blood Pressure
- **SICU** - Surgical Intensive Care Unit
- **SQI** - Signal Quality Index
- **SPO2** - Peripheral Oxygen Saturation
- **TCO** - Thermodilution Cardiac Output
- **T-SICU** - Trauma Surgical Intensive Care Unit
- **WFDB** - Waveform Database software package – See http://www.physionet.org/

Chapter 6

Appendix

6.1 Database Schema

The full database schema documentation is available on the MIMIC II website:
http://mimic.physionet.org/schema/latest

6.2 Multiple ID mappings

Except for "invasive diastolic arterial blood pressure", all parameter sample values are in the "value1num" column in the CHARTEVENTS table. The systolic and diastolic invasive arterial blood pressure are stored under the same ITEMID; with the systolic value in "value1num", and diastolic in "value2num".

6.2.1 Bicarbonate (HCO3)

```
ITEMID  LABEL      CATEGORY
-------------------
787     Carbon Dioxide    Chemistry
812 HCO3              [value1num]
3810 Total CO2
```

6.2.2 Bilirubin (highest)

```
ITEMID  LABEL      CATEGORY
-------------------
848     Total Bili (0-1.5)    Chemistry
1538    Total Bili            Chemistry
```

6.2.3 Blood Pressure

Invasive (Arterial) Blood Pressure (IABP/IBP)

```
ITEMID  LABEL      CATEGORY
-------------------
52 Arterial Blood Pressure (Mean)
51 Arterial Blood Pressure (Systolic)  [value1num]
51 Arterial Blood Pressure (Diastolic) [value2num]
```

Invasive blood pressures are generally more accurate than non-invasive blood pressures.

Non Invasive Blood Pressure (NIBP)

```
ITEMID  LABEL      CATEGORY
-------------------
455     nabpsys  value1num
455     nabpdias  value2num
456     nabpmean  value1num
1149    NBP:
751     zzzNBP
```

6.2.4 Blood Transfusions

```
ITEMID  LABEL      CATEGORY
-------------------
  734 Packed RBC's 350.0ml
  31 RBC'S
  144 Packed RBC's
  172 OR Packed RBC's
  397 Washed PRBC's
  980 Packed RBC's 150.0ml
  1011 Packed RBC's 750.0ml
  1106 Packed RBC's 75.0ml
  1141 Packed RBC's 372.0ml
  1585 Packed RBC's 100.0ml
  1737 Packed RBC's 125.0ml
  1738 Packed RBC's 400.0ml
  2909 Packed RBC's 325.0ml
  3735 Packed RBC's 95.0ml
  3992 Packed RBC's 500.0ml
  4245 Packed RBC's 3.0ml
  4258 Packed RBC's 450.0ml
  4422 Packed RBC's 300.0ml
```

6.2.5 Cardiac Output (CO)

```
ITEMID  LABEL      CATEGORY
-------------------
  90 C.O.(thermodilution)
  89 C.O. (fick)
  1601 C.C.O
  2112 continuous C.O
```

Note that thermodilution CO calculations are generally more accurate than those calculated through the Fick method.

6.2.6 Carbon Dioxide (CO2)

```
ITEMID  LABEL      CATEGORY
-------------------
777     ArtCO2Calc    [value1num]
4199    ArtCO2Calc    [value1num]
787     CarbonDioxide [value1num]
3808    CarbonDioxide [value1num]
3810    CarbonDioxide [value1num]
```

6.2.7 Creatinine (highest)

```
ITEMID  LABEL      CATEGORY
-------------------
791 Creatinine (0-1.3)
3750 Creatinine   (0-0.7)
1525 Creatinine
```

6.2.8 Central Venous Pressure (CVP)

```
ITEMID  LABEL      CATEGORY
-------------------
  1103 cvp
  113 CVP
```

6.2.9 Glucose Levels

```
ITEMID  LABEL      CATEGORY
-------------------
811 Glucose (70-105)    [value1num]
1529 Glucose             [value1num]
```

6.2.10 Intra-aortic balloon (IABP) pump rate

```
ITEMID  LABEL      CATEGORY
-------------------
  224 IABP Mean
  225 IABP setting
  2162 IABP TIMING ADJUSTED
  2391 IABP
  2515 IABP-BP
  2865 iabp-bp
  6424 IABP BP
```

6.2.11 Intra-cranial Pressure (ICP)

```
ITEMID  LABEL      CATEGORY
-------------------
  1374 ICP Right
  226 ICP
  2045 icp left
  2745 ICP LEFT
  5856 icp
```

6.2.12 IV Fluids

```
ITEMID  LABEL      CATEGORY
-------------------
1       HourlyIn           pervolume
2       HourlyOut          pervolume
2       DailyOut           cumvolume
18      HourlyIV           pervolume
29      NetHourlyBalance   pervolume
```

6.2.13 Lactate

```
ITEMID  LABEL      CATEGORY
-------------------
818 Lactic Acid(0.5-2.0)
1531 Lactic Acid
```

6.2.14 Oxygen Saturation (SpO2/SaO2)

```
ITEMID  LABEL      CATEGORY
-------------------
  1148 SpO2:
  646 SpO2
  834 SaO2
```

6.2.15 pH

```
ITEMID  LABEL      CATEGORY
-------------------
865     pH value1num
1126    pH value1num
780     pH value1num
4202    pH value1num
4753    pH value1num
```

6.2.16 Potassium

```
ITEMID  LABEL      CATEGORY
-------------------
829 Potassium (3.5-5.3) [value1num]
1535 Potassium           [value1num]
3792 Potassium  (3.5-5.3) [value1num]
```

6.2.17 Pressor Medications

```
ITEMID  LABEL      CATEGORY
-------------------
 46     Isuprel
 47 Levophed
 120 Levophed-k
 43 Dopamine
 307 Dopamine Drip
 44 Epinephrine
 119 Epinephrine-k
 309 Epinephrine Drip
 51 Vasopressin
 127 Neosynephrine
 128 Neosynephrine-k
```

Note: Drugs interact, and can often have the oppositcve effect you might expect when the drug dose is low or high. For example, the vasodepressor action of Isuprel is reversed to vasopressor action by small doses of ergotamine or ergotoxine. This reversal is associated with a marked increase in the amplitude of ventricular contraction, in pulse pressure and in rate Lands et al. [1950].

6.2.18 Pulmonary Arterial Pressure (PAP)

```
ITEMID  LABEL      CATEGORY
--------------------------
  491 PAP Mean
  492 PAP S/D
```

6.2.19 Respiration Rate

```
ITEMID  LABEL      CATEGORY
--------------------------
  614 Resp Rate (Spont)
  615 Resp Rate (Total)
  618 Respiratory Rate  (*)
  653 Spont. Resp. Rate
  1151 Respiratory Rate:
  1635 HIGH Resp Rate
  1884 Spont Resp rate
  2117 Low resp rate
  3603 Resp Rate
  3337 Breath Rate
```

(*) indicates preferred ITEMID.

6.2.20 Sodium

```
ITEMID  LABEL      CATEGORY
--------------------------
837 Sodium (135-148)   [value1num]
1536 Sodium             [value1num]
3803 Sodium  (135-148)  [value1num]
```

6.2.21 Temperature

```
ITEMID  LABEL      CATEGORY
--------------------------
676 Temperature C          [value1num]
677 Temperature C (calc)   [value1num]
678 Temperature F          [value1num]
679 Temperature F (calc)   [value1num]
```

6.2.22 Urine Output

```
ITEMID  LABEL      CATEGORY
--------------------------
26 Urine Out Total
3053 URINE OUT
3165 Urine Output Total
3175 Urine
3462 urine
3519 urine amnt
```

6.2.23 Ventilators

```
ITEMID  LABEL      CATEGORY
--------------------------
505     peep              [value1num]
506     peep              [value1num]
535     PeakInspPressure  [value1num]
543     PlateauPressure   [value1num]
544     Plateau Time (7200)
545     Plateau-Off
619     Respiratory Rate Set
 39     Airway Size
535     Peak Insp. Pressure
683     Tidal Volume (Set)
720     Ventilator Mode
721     Ventilator No.
722     Ventilator Type
732     Waveform-Vent
```

6.2.24 White Blood Cell Count (WBC)

```
ITEMID  LABEL      CATEGORY
--------------------------
1542 WBC                 [value1num]
1127 WBC   (4-11,000)    [value1num]
861 WBC (4-11,000)      [value1num]
4200 WBC 4.0-11.0        [value1num]
```

6.3 Commonly used parameters

Together with the above parameters in section 6.2, the following list may be helpful.

```
ITEMID  LABEL      CATEGORY
--------------------------
770     AST         [value1num]
781 BUN (6-20)  [value1num]
198 GCS         [Total]
828     Platelets   [value1num]
211     heartrate   [value1num]
813     Hematocrit  [value1num]
20001   SAPS1       [value1num]
504  PCWP        [value1num]
```

GCS = Glasgow Coma Scale. BUN = Blood Urea Nitrogen (also known as Urea or Urea nitrogen). SAPS1 indicates Simplified Acuity Score (version 1). PCWP = Pulmonary Capillary Wedge Pressure (or simply 'Wedge Pressure').

6.4 Frequency of all ICD-9 codes for adult ICU-related hospital admissions

Table 6.1 lists the most frequent ICD-9 codes (including the primary ICD-9 codes) for the same population for the thirty-two major ICD-9 categories. Note that there is one extra category in the non-primary code categories (see table 6.1) "Supplementary classification of external causes of injury and poisoning" (code range E800 to E999). These ICD-9 codes are not used for primary classification.

Table 6.1: Distribution of major categories of ICD-9 codes for adult ICU-related hospital admissions (n = 211,416)

Category	**Code Range**	**Number of Codes**	**%**
Metabolic disorder	240-279	23927	11.32%
Pulmonary disease	460-519	18694	8.84%
Other forms of heart disease	420-429	13203	6.25%
Digestive disease	520-579	13181	6.23%
Supplementary classification of factors influencing health status and contact with health services	V01-V86	12801	6.05%
Ischemic heart disease	410 - 414	12794	6.05%
Renal insufficiency	580-629	11262	5.33%
Hypertensive disease	401-405	10970	5.19%
Symptoms, signs, and ill-defined conditions	780-799	8777	4.15%
Diseases of the blood and blood-forming organs	280-289	8744	4.14%
Trauma	800-959	7975	3.77%
Heart failure	428	7470	3.53%
Infectious diseases	001-139	7388	3.49%
Mental disorders	290-319	7337	3.47%
Supplementary classification of external causes of injury and poisoning	E800-E999	6632	3.14%
Arteries and veins	440-459	5828	2.76%
Neoplasms	140-239	5403	2.56%
Neurologic disease	320-389	5140	2.43%
Diseases of the musculoskeletal system & connective tissue	710-739	3632	1.72%
Other complications of procedures, NEC	998	2936	1.39%

continued on next page

continued from previous page			
Category	**Code Range**	**Number of Codes**	**%**
Cerebrovascular disease	430-438	2849	1.35%
Diseases of the skin and subcutaneous tissue	680-709	2799	1.32%
Complications affecting specified body systems, not elsewhere classified	997	2551	1.21%
Complications peculiar to certain specified procedures	996	2403	1.14%
Other and unspecified effects of external causes	990-995	1941	0.92%
Chronic rheumatic heart disease	393-398	1438	0.68%
Diseases of pulmonary circulation	415-417	1233	0.58%
Congenital anomalies	740-759	673	0.32%
Complications of pregnancy, childbirth, and the puerperium	630-677	627	0.30%
Poisoning	960-989	590	0.28%
Complications of medical care, not elsewhere classified	999	213	0.10%
Acute Rheumatic fever	390-392	5	0.00%
Total		**211416**	**100.00%**

Bibliography

Christine L. Tsien and James C. Fackler. An annotated data collection system to support intelligent analysis of intensive care unit data. In *IDA '97: Proceedings of the Second International Symposium on Advances in Intelligent Data Analysis, Reasoning about Data*, pages 111–121, London, UK, 1997. Springer-Verlag. ISBN 3-540-63346-4. URL http://www.springerlink.com/content/n89atcgqklca11v2/.

Patrick R. Norris and Benoit M. Dawant. Closing the loop in ICU decision support: Physiologic event detection, alerts, and documentation. *J Am Med Inform Assoc*, 9(90061):S102–107, 2002. doi: 10.1197/jamia.M1238. URL http://www.jamia.org/cgi/content/abstract/9/6_suppl_1/S102.

R. R. Abbott, M. Setter, S. Chan, and K. Choi. APACHE II: prediction of outcome of 451 ICU oncology admissions in a community hospital. *Ann Oncol*, 2(8):571–574, Sep 1991. URL http://annonc.oxfordjournals.org/content/2/8/571.short.

Laurent G Glance, Turner M Osler, and Andrew W Dick. Identifying quality outliers in a large, multiple-institution database by using customized versions of the Simplified Acute Physiology Score II and the Mortality Probability Model II0. *Crit Care Med*, 30(9):1995–2002, Sep 2002. doi: 10.1097/01.CCM.0000026324.64324.59. URL http://dx.doi.org/10.1097/01.CCM.0000026324.64324.59.

G. D. Clifford, W. J. Long, G. B. Moody, and P. Szolovits. Robust parameter extraction for decision support using multimodal intensive care data. *Phil Trans Royal Soc A*, 367(1877):411–429, January 2009. doi: doi:10.1098/rsta.2008.0157. URL http://www.ncbi.nlm.nih.gov/pubmed/18936019. Special issue on Signal Processing in Vital Rhythms and Signs.

I. Neamatullah, M. Douglass, L. H. Lehman, A. Reisner, M. Villarroel, W. J. Long, P. Szolovits, G. B. Moody, R. G. Mark, and G. D. Clifford. Automated de-identification of free-text medical records. *BMC Med Inform Decis Mak*, 8(32), 2008. doi: doi:10.1186/1472-6947-8-32. URL http://www.biomedcentral.com/1472--6947/8/32/.

W. Zong, G. B. Moody, and R. G. Mark. Reduction of false arterial blood pressure alarms using signal quality assessment and relationships between the electrocardiogram and arterial blood pressure. *Med Biol Eng Comput*, 42:698–706, 2004. URL http://mimic.physionet.org/Archive/Publications/ZongMBEC04.pdf.

J. X. Sun, A. T. Reisner, and R. G. Mark. A signal abnormality index for arterial blood pressure waveforms. *Comput Cardiol*, 33:13–16, September 2006. URL http://www.cinc.org/Proceedings/2006/pdf/0013.pdf.

Q. Li, R. G. Mark, and G. D. Clifford. Robust heart rate estimation from multiple asynchronous noisy sources using signal quality indices and a Kalman filter. *IOP Physiol Meas*, 29(1):15–32, January 2008. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2259026/. (Awarded the Martin Black Prize for Best Paper in Physiological Measurement in 2008).

Q. Li, R. G. Mark, and G. D. Clifford. Artificial arterial blood pressure artifact models and an evaluation of a robust blood pressure and heart rate estimator. *BMC Biomed Eng Online*, 8(13), 2009. doi: doi:10.1186/1475-925X-8-13. URL http://www.biomedical-engineering-online.com/content/8/1/13.

A. Aboukhalil, L. Nielsen, M. Saeed, R. G. Mark, and G. D. Clifford. Reducing false alarm rates for critical arrhythmias using the arterial blood pressure waveform. *J Biomed Inform*, 41(3):442–451, June 2008. ISSN 1532-0464. doi: http://dx.doi.org/10.1016/j.jbi.2008.03.003. URL http://www.ncbi.nlm.nih.gov/pubmed/18440873.

M. Saeed, G. D. Clifford, M. Villarroel, G. B. Moody, L. Lehman, O. Abdala, M. M. Douglass, J. Frassica, T. Heldt, C. Hug, B. A. Janz, T. H. Kyaw, C. Lieu, W. J. Long, B. Moody, L. Nielsen, I. Neamatullah, G. Raber, A. Reisner, P. Szolovits, G. Verghese, and R. G. Mark. MIMIC-II: A major new database to support research in ICU patient monitoring and clinical decision support systems. *Critical Care*, 2009. In Submission.

Alan Beaulieu. *Learning SQL*. O'Reilly, August 2005. ISBN 0-596-00727-2. URL http://books.google.com/books?id=1PgCCVryjOQC&printsec=frontcover&dq=Beaulieu+A.+++++Learning+SQL&hl=en&ei=Aal2TI3SC4H6lwewkaXtCw&sa=X&oi=book_result&ct=result&resnum=1&ved=0CDUQ6AEwAA.

A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. Ch. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C. K. Peng, and H. E. Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000. URL http://circ.ahajournals.org/cgi/content/full/101/23/e215.

G.B. Moody and R.G. Mark. A database to support development and evaluation of intelligent intensive care monitoring. *Computers in Cardiology 1996*, pages

657–660, Sep 1996. doi: 10.1109/CIC.1996.542622. URL http://ecg.mit.edu/george/publications/mimic-cinc-1996.pdf.

Java Database Connectivity (JDBC). URL http://java.sun.com/products/jdbc/overview.html.

O. T. Abdala and M. Saeed. Estimation of missing values in clinical laboratory measurements of ICU patients using a weighted K-nearest neighbors algorithm. *Comput Cardiol*, 31:693–696, 2004. URL http://www.cinc.org/Proceedings/2004/pdf/693.pdf.

C. Hug and G. D. Clifford. An analysis of the errors in recorded heart rate and blood pressure in the ICU using a complex set of signal quality metrics. *Comput Cardiol*, 34:641–645, September 2007. URL http://www.cinc.org/Proceedings/2007/pdf/0641.pdf.

A. M. Lands, F. P. Luduena, J. I. Grant, Estelle Ananenko, and M. L. Tainter. Reversal of the depressor action of n-isopropylarterenol (isuprel) by ergotamine and ergotoxine. *J Pharmacol Exp Ther*, 100(3):284–297, 1950. URL http://jpet.aspetjournals.org/cgi/content/abstract/100/3/284.