Multiparameter Intelligent Monitoring in Intensive Care II: A public-access intensive care unit database*

Mohammed Saeed, MD, PhD; Mauricio Villarroel, MBA; Andrew T. Reisner, MD; Gari Clifford, PhD; Li-Wei Lehman, PhD; George Moody; Thomas Heldt, PhD; Tin H. Kyaw, MEng; Benjamin Moody; Roger G. Mark, MD, PhD

Objective: We sought to develop an intensive care unit research database applying automated techniques to aggregate high-resolution diagnostic and therapeutic data from a large, diverse population of adult intensive care unit patients. This freely available database is intended to support epidemiologic research in critical care medicine and serve as a resource to evaluate new clinical decision support and monitoring algorithms.

Design: Data collection and retrospective analysis.

Setting: All adult intensive care units (medical intensive care unit, surgical intensive care unit, cardiac care unit, cardiac surgery recovery unit) at a tertiary care hospital.

Patients: Adult patients admitted to intensive care units between 2001 and 2007.

Interventions: None.

Measurements and Main Results: The Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II) database consists of 25,328 intensive care unit stays. The investigators collected detailed information about intensive care unit patient stays, including laboratory data, therapeutic intervention profiles such as vasoactive medication drip rates and ventilator settings, nursing progress notes, discharge summaries, radiology reports, provider order entry data, International Classification of Diseases, 9th Revision codes, and, for a subset of patients, high-resolution vital sign trends and waveforms. Data were automatically deidentified to comply with Health Insurance Portability and Accountability Act standards and integrated with relational database software to create electronic intensive care unit records for each patient stay. The data were made freely available in February 2010 through the Internet along with a detailed user's guide and an assortment of data processing tools. The overall hospital mortality rate was 11.7%, which varied by critical care unit. The median intensive care unit length of stay was 2.2 days (interquartile range, 1.1–4.4 days). According to the primary International Classification of Diseases, 9th Revision codes, the following disease categories each comprised at least 5% of the case records: diseases of the circulatory system (39.1%); trauma (10.2%); diseases of the digestive system (9.7%); pulmonary diseases (9.0%); infectious diseases (7.0%); and neoplasms (6.8%).

Conclusions: MIMIC-II documents a diverse and very large population of intensive care unit patient stays and contains comprehensive and detailed clinical data, including physiological waveforms and minute-by-minute trends for a subset of records. It establishes a new public-access resource for critical care research, supporting a diverse range of analytic studies spanning epidemiology, clinical decision-rule development, and electronic tool development. (Crit Care Med 2011; 39:952–960)

KEY WORDS: databases; clinical decision support; hemodynamic instability; information technology; patient monitoring

W e report the establishment of the Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II) research database that is notable for four factors: it is publicly and freely available to other research organizations upon request; it encompasses a diverse population of intensive care unit (ICU) patients; it contains high temporal resolution data, including laboratory results, electronic clinical documentation, and bedside monitor numeric trends and waveforms (such as the electrocardiogram); and it has been deidentified in a Health Insurance Portability and Accountability Act-compliant manner. The MIMIC-II database will support a diverse range of analytic studies spanning epidemiology, clinical decision-rule development, and electronic tool development.

Historically, large-scale ICU databases have been effective resources to understand risk factors and natural histories of critical illness as well as the efficacy of various treatment strategies. For instance, Acute Physiology and Chronic Health Evaluation I–III and Project Impact contained daily abstractions of patient data that provided new insights and

scoring tools to relate patient outcomes and lengths of stays with the patients' conditions on admission (1, 2). Such collection and analysis of large volumes of ICU data are invaluable to the advancement of clinical knowledge, but it is extremely effort-intensive because there are substantial challenges to the collection of the data. Such difficulties include: disparate sources of data, eg, clinical documentation versus laboratory results; erroneous or missing data; unsynchronized time references; proprietary data formats; limitations of computing power, networking bandwidth, and digital storage capacity; and concerns related to patient privacy. The challenge of data collection has sometimes been addressed through coordinated efforts by a network of clinical investigators interested in specific problem domains such as acute respiratory distress syndrome (ARDSNET Trial) (3), acute kidney injury (4), or septic shock (5). However, these powerful disease specific databases were not designed to be exploited as research resources to support other domains of ICU research nor are their data widely available.

In 2003, under National Institutes of Health funding, we established a research program with the objective of developing and evaluating advanced ICU monitoring and decision support systems. A critical requirement of our program was the development of a substantial and comprehensive clinical database from ICU patients. Now, 7 yrs later, the MIMIC-II database has reached a state of maturity sufficient to be made available to the wider research community. The database is intended to support a wide diversity of research in critical care. Unlike related databases, there are no access fees or extensive credentialing requirements, and documentation and other support are available so that the data will be accessible to the largest community of researchers.

This article contains a detailed report of the MIMIC-II data acquisition process, which was accomplished through collaboration among academic, industrial, and clinical groups. Summary statistics are provided to characterize the database and we provide examples of clinical hypotheses and physiologic signal processing algorithms we have studied with MIMIC-II. The high temporal resolution parameters within the database such as hourly vital sign trends, ventilator settings, intravenous medication drip rates, and fluid balances enable novel investigations of transient clinical outcomes such as

Table 1. Description of clinical data classes[a]

Clinical Data Class	Description
General	Patient demographics, hospital admission and discharge dates, room tracking, code status, hospital death dates (in or out of the ICU), ICD-9 codes, etc
Physiological	Hourly nurse-verified vital signs (BP, HR, etc), SAPS, ventilator settings, etc
Clinical laboratory tests	Hematology, blood chemistries, ABGs, urinalysis, microbiology, etc
Medications	Detailed administration records of IV medications, provider order entry data
Fluid balance	Hourly and cumulative intake (solutions, blood, etc) and output (urine, estimated blood loss, etc)
Reports	Free text reports of imaging studies (x-ray, CT, MRI), 12-lead ECGs, echocardiograms, etc
Notes	Free text notes including nursing and respiratory therapist progress notes; physician hospital discharge summaries

ICU, intensive care unit; ICD-9, International Classification of Diseases, 9th Revision; BP, blood pressure; HR, heart rate; SAPS, Simplified Acute Physiological Score; ABGs, arterial blood gases; IV, intravenous; CT, computed tomography; MRI, magnetic resonance imaging; ECGs, electrocardiograms.

[a]A comprehensive listing of the hundreds of clinical parameters available in the MIMIC-II database is available at http://physionet.org/mimic2.

hypotensive episodes. Similarly, MIMIC-II enables the analysis of transient independent variables such as electrocardiogram waveform features and their associated clinical outcomes. The unique features of MIMIC-II are compared with other major databases and we discuss the major challenges encountered in developing MIMIC-II and explore future improvements. The MIMIC-II database takes advantage of improvements in healthcare information technologies to establish a new standard in public-access databases for critical care research.

MATERIALS AND METHODS

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 because the study did not impact clinical care and all protected health information was deidentified.

Patient Population. This first release of the MIMIC-II database encompasses virtually all adult patients admitted to ICUs at Boston's Beth Israel Deaconess Medical Center during the period 2001–2007; additional MIMIC-II data collection is ongoing. Boston's Beth Israel Deaconess Medical Center is a 620-bed tertiary academic medical center in Boston and a level I trauma center with 77 critical care beds. The ICUs are closed with 24-hr inhouse intensivist supervision of patient care. These ICUs include the medical, surgical, coronary, and cardiac surgery recovery care units. ICU stays separated by >24 hrs were counted separately even if they occurred within the same hospital stay.

Database Development. The data acquisition process was not visible to staff and did not interfere with the clinical care of patients or methods of monitoring. Two categories of data were collected: clinical data, which were aggregated from ICU information systems and hospital archives, and high-resolution physiological data (waveforms and time series of derived physiological measurements) that were obtained from bedside monitors.

Clinical Data. Clinical data were obtained from the CareVue Clinical Information System (models M2331A and M1215A; Philips Healthcare, Andover, MA) deployed in all the study ICUs as well as from hospital electronic archives (Table 1). The data included such items as time-stamped nurse-verified physiological measurements (eg, hourly documentation of heart rate, arterial blood pressure, pulmonary artery pressure, etc); nurses' and respiratory therapists' progress notes; continuous intravenous drip medications; fluid balances; patient demographics; interpretations of imaging studies; physician orders; discharge summaries; and International Classification of Diseases, 9th Revision (ICD-9) codes. Comprehensive diagnostic laboratory results (eg, blood chemistries, complete blood counts, arterial blood gases, microbiology results) were obtained from the patients' entire hospital stay, including periods outside the ICU.

Physiological Data. Physiological data were obtained with the technical assistance of the monitoring system vendor. Patient monitors (Component Monitoring System Intellivue MP-70; Philips Healthcare) were located by every ICU patient bed. Each monitor acquired and digitized multiparameter physiological data; processed the signals to derive time series (trends) of clinical measures such as heart rate, blood pressures, and oxygen saturation, etc; and also produced bedside mon-

Table 2. MIMIC-II patient population by critical care unit

Critical Care Unit	MICU	SICU	CSRU	CCU	Total
Hospital admissions, no. (% of total admissions)	8,700 (38.0%)	6,004 (26.3%)	4,707 (20.6%)	3,459 (15.1%)	22,870 (100%)
Distinct ICU stays, no.[a] (% of total unit stays)	9,683 (38.2%)	6,730 (26.6%)	5,060 (20.0%)	3,855 (15.2%)	25,328 (100%)
Waveform records, no. (% of unit stays)	1,662 (17.1%)	524 (7.8%)	710 (14.0%)	971 (25.2%)	3,867 (15.3%)
Records with matched waveforms, no. (% of unit stays)	794 (9.9%)	200 (5.5%)	390 (4.8%)	677 (12.2%)	2,061 (8.1%)
Age, yrs, mean ± SD	63.0 ± 18.3	59.3 ± 19.6	65.4 ± 13.7	68.6 ± 15.1	63.3 ± 17.7
Gender, male, percent of unit stays	49.9%	57.4%	65.4%	58.3%	56.3%
ICU length of stay, median days (IQR)	2.2 (1.1–4.5)	2.4 (1.2–5.5)	2.2 (1.1–4.1)	1.9 (1.0–3.6)	2.2 (1.1–4.4)
Hospital length of stay, median days (IQR)	7.0 (4–13)	8 (5–16)	8 (5–12)	5 (3–9)	7 (4–13)
SAPS I score, day 1, median (IQR)[b]	13 (10–17)	13 (9–17)	15 (12–18)	11 (8–15)	13 (10–17)
SAPS I score, day 2, median (IQR)[b]	12 (9–15)	12 (9–15)	11 (8–13)	11 (7–14)	11 (9–15)
SAPS I score, day 3, median (IQR)[b]	13 (10–16)	12 (9–15)	11 (8–14)	11 (8–14)	12 (9–15)
Mechanical ventilation, no. (% of unit stays)	3,456 (35.7%)	3,418 (50.8%)	4,165 (82.3%)	930 (24.1%)	11,969 (47.3%)
Invasive Swan-Ganz hemodynamic monitoring, no. (% of unit stays)	348 (3.6%)	1,025 (15.2%)	3,278 (64.7%)	989 (25.7%)	5,637 (22.3%)
Invasive arterial blood pressure monitoring, no. (% of unit stays)	3,283 (33.9%)	4,502 (66.9%)	4,503 (89.0%)	1,766 (45.8%)	14,054 (55.5%)
Use of vasoactive medications, no. (% of unit stays)	2,356 (24.3%)	1,655 (24.6%)	3,549 (70.1%)	1,133 (29.4%)	8,693 (34.3%)
ICU mortality, percent of unit stays	14.5%	11.0%	3.3%	9.0%	10.5%
Hospital mortality, percent of unit stays	16.2%	12.4%	3.6%	10.0%	11.7%

MIMIC-II, Multiparameter Intelligent Monitoring in Intensive Care II; MICU, medical intensive care unit; SICU, surgical intensive care unit; CSRU, cardiac surgery recovery unit; CCU, coronary care unit; ICU, intensive care unit; IQR, interquartile range; SAPS, Simplified Acute Physiological Score.

[a]Multiple ICU stays during the same hospitalization that are separated by greater than 24 hrs are considered separately. Thus, it is possible for one hospital admission to encompass more than one ICU stay; [b]excludes 20.5% of overall ICU stays without all parameters necessary for SAPS I calculation.

itor alarms. Those data were all transmitted to a networked nursing central station within each ICU (M3155 Intellivue Information Center; Philips Healthcare). The physiological waveforms (such as electrocardiogram, blood pressures, pulse plethysmograms, respirations) were sampled at 125 Hz, and trend data were updated each minute. The data were subsequently stored temporarily in a central database server that typically supported several ICUs. A customized archiving agent, developed through collaboration with Philips Healthcare, created permanent copies of the physiological data residing in central database servers. The data were physically transported from the hospital to the laboratory every 2 to 4 wks where they were deidentified, converted to an open source data format (6), and incorporated into the MIMIC II waveform database. Limited capacity and intermittent failures of the archiving agents restricted waveform collection to a fraction (15%) of the monitored ICU beds (Table 2). No attempt was made to assure that the ICU records with waveform/trend data were statistically representative of the database as a whole.

Database Merger and Postprocessing. The second stage in developing the MIMIC-II database involved significant data postprocessing and database organization to obtain integrated medical records for each patient. Across the hospital's clinical databases, patients are identified by their unique Medical Record Numbers and their Encounter Numbers (the latter uniquely identifies a particular hospitalization for patients who might have been admitted multiple times), which we relied on in merging information from different hospital sources. Matching waveform records to clinical data was based on unique identifiers such as medical record numbers, dates of admission, and patient names. Sometimes, however, nurses did not enter patient identifiers into the bedside monitors and as a result, only approximately half of the available waveform records could be uniquely matched to clinical data. More information on database merger, in particular, how database integrity was ensured, is available at the MIMIC-II web site (http://physionet.org/mimic2).

An additional task was to convert the patient monitoring data from Philips' proprietary format into an open-source format. With assistance from the medical equipment vendor, the waveforms, trends, and alarms were translated into WFDB, an open data format that is used for publicly available databases on the National Institutes of Health-sponsored *PhysioNet* web site (6).

Because MIMIC-II is intended to be a reflection of real-life clinical data (rather than pristine data derived from a carefully conducted clinical trial), the clinical and physiological content of the database has not been altered. In other words, we did not enforce any range restrictions or other sanity checks on the data (beyond what each individual hospital database might impose).

Deidentification and Health Insurance Portability and Accountability Act Compliance. All data that were integrated into the MIMIC-II database were deidentified in compliance with Health Insurance Portability and Accountability Act standards to facilitate public access to MIMIC-II. Deletion of protected health information from structured data sources was straightforward (eg, database fields that provide the patient name, date of birth, etc). As well, we removed protected health information from the discharge summaries, diagnostic reports, and the approximately 700,000 free-text nursing and respiratory notes in MIMIC-II using an automated algorithm that has been shown to have superior performance in comparison to clinicians in detecting protected health information (7). This algorithm accommodates the broad spectrum of writing styles in our data set, including personal variations in syntax, abbreviations, and spelling. We have posted this algorithm in open-source form as a general tool to be used by others for deidentification of free-text notes (8).

Database Distribution and Documentation. The MIMIC-II database was developed with the intention to make its contents freely available to interested researchers. The Internet is an effective distribution mechanism to facilitate the dissemination of the deidentified MIMIC-II database. To restrict traffic to legitimate medical researchers, access requires completion of a simple data use agreement and proof that the researcher has completed human subjects training.

The MIMIC-II database is available in two forms. In the first form, interested researchers can obtain a flat-file text version of the clinical database and the associated database schema that enables them to reconstruct the database using their method of choice. In the second form, interested researchers can gain access to the database through a password-protected web service. Database searches require the users to familiarize themselves with the database layouts and to program database queries using

the Structured Query Language. Query output can be exported to comma-separated files to be analyzed offline using statistical or other software. Accessing and processing data from MIMIC-II is complex. It is highly recommended that studies based on the MIMIC-II database be conducted as collaborative efforts that include clinical, statistical, and relational database expertise.

Detailed documentation and procedures for obtaining access to MIMIC-II are available at the MIMIC-II web site (http://physionet.org/mimic2).

Database Characterization. We report the characteristics of version 2.4 of the MIMIC-II database (released on February 1, 2010) so that investigators can determine whether a potential study of interest is possible with MIMIC-II. Clinical data that are summarized include mortality, length of stay in the ICU, primary ICD-9 codes, patient demographics, and frequency of use of significant therapeutic interventions. The source code used to generate the statistical results provided in this article is publicly available (9).

Acuity scores were not routinely documented for MIMIC-II patients during the admission process. We implemented an automated algorithm to retrospectively compute the ICU Simplified Acute Physiological Score (SAPS) I scores for the first 3 days of all admissions with complete SAPS I data. The SAPS I formula was chosen for its simplicity, requiring only available clinical laboratory measurements, fluid balance, and vital signs (There have been several refinements to the original SAPS algorithm that incorporate the presence or absence of chronic disease (such as cancer, AIDS, etc) in the overall acuity of a patient. Such clinical data exist in MIMIC-II free-text discharge summaries.). The mortality rate was trended as a function of admission SAPS I scores as a method to validate the automated SAPS I calculations. The distribution of problems, mortality rates, and acuity was also analyzed across the different adult ICUs (coronary care unit, medical ICU, surgical ICU, cardiac surgery recovery unit).

To illustrate some of the analyses possible with the database, we computed the incidence, associated mortality, and odds ratios of a range of diseases that can be defined by objective laboratory abnormalities and of vital sign abnormalities, including heart rate, systolic blood pressure, respiratory rate, and arterial oxygen saturation. We only included the population in whom those diagnostic data were ordered. We excluded the small population with multiple ICU stays in case their characteristics were unrepresentative, and we excluded the final 12 hrs of any ICU stay to avoid confounding the results with physiology associated with withdrawal of care.

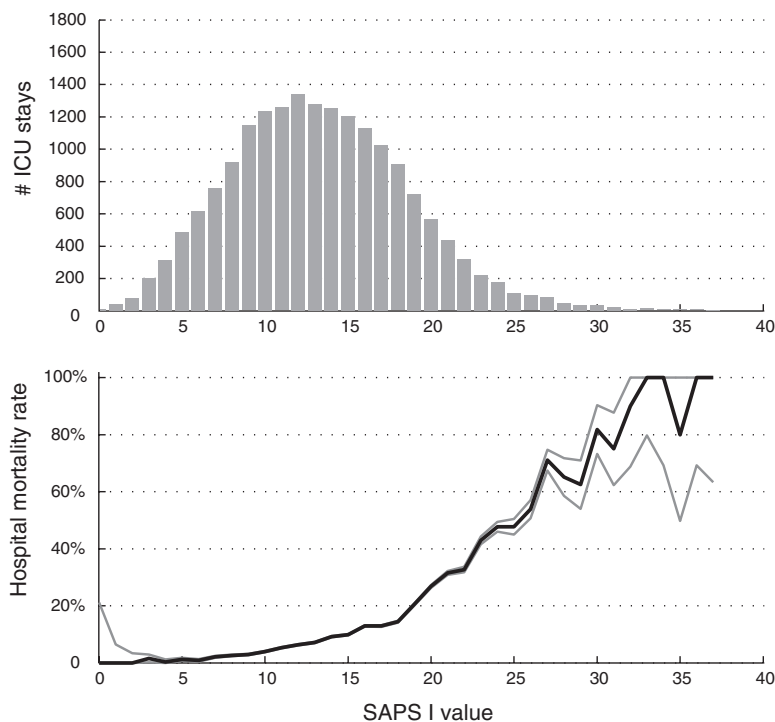Lastly, we compiled a comparison of MIMIC-II to other ICU research databases.

Figure 1. Histogram of admission Simplified Acute Physiological Score (*SAPS*) I values for Multiparameter Intelligent Monitoring in Intensive Care II patients (*top panel*) and associated mortality (*bottom panel*) with 95% confidence intervals.

RESULTS

Table 2 includes summary statistics and patient demographics across the major adult ICU patient populations in MIMIC-II (medical ICU, surgical ICU, cardiac surgery recovery unit, coronary care unit). The database (version 2.4) encompasses 25,328 ICU stays from 22,870 hospital admissions. Of those admissions, 1360 (5.9%) had multiple ICU stays, and on average there were 1.11 ICU stays per hospitalization. The median (interquartile range) ICU stay lasted 2.2 days (1.1–4.4), whereas the median hospital length of stay was 7 days (4–13). The median (interquartile range) ICU length of stay was longest in the surgical ICU, 2.4 days (1.2–5.5) and shortest in the coronary care unit, 1.9 days (1.0–3.6).

The overall hospital mortality rate in the MIMIC-II database was 11.7%. The mortality rate in MIMIC-II patients is trended as a function of SAPS I score in Figure 1. As the admission SAPS I score increased, the mortality rate significantly increased. There were differences in the SAPS I scores and the mortality rates among the different care units. The postoperative patient population in the cardiac surgical recovery unit tended to have a higher SAPS I scores on day 1 (reflecting intubation and sedation), which be-

came comparable to the other units by day 2. The medical ICU had the highest hospital mortality rate (16.2%) (chi-squared vs. other units, $p < .001$), whereas the cardiac surgery recovery unit had the lowest hospital mortality rate (3.6%) ($p < .001$).

Table 3 provides the primary ICD-9 codes from the patients' hospital discharges. The following disease categories each comprised at least 5% of the discharge codes: diseases of the circulatory system (39.1%); trauma (10.2%); diseases of the digestive system (9.7%); pulmonary diseases (9.0%); infectious diseases (7.0%); and neoplasms (6.8%). Table 4 reports the incidence and associated mortality of a range of diseases that can be defined by objective laboratory abnormalities. Table 5 reports the incidence and associated mortality of vital sign abnormalities, including heart rate, systolic blood pressure, respiratory rate, and arterial oxygen saturation.

Comparison With Other Databases. Table 6 provides a comparison of MIMIC-II with other databases. The salient database features that are compared include the size of the databases (in terms of number of ICU records), record completeness (availability of different physiological and clinical data), and availability.

Table 3. Distribution of primary ICD-9 codes in MIMIC-II (version 2.4)

	MICU Stays, No. (% of Unit Stays)	SICU, No., (% of Unit Stays)	CSRU, No., (% of Unit Stays)	CCU, No. (% of Unit Stays)	Total No. (% of Unit Stays)
Infectious and parasitic diseases, ie, septicemia, other infectious and parasitic diseases, etc (001–139)	1,319 (15.2%)	179 (3.0%)	30 (0.6%)	75 (2.2%)	1,603 (7.0%)
Neoplasms of digestive organs and intrathoracic organs, etc (140–239)	651 (7.5%)	630 (10.5%)	212 (4.5%)	69 (2.0%)	1,562 (6.8%)
Endocrine, nutritional, metabolic, and immunity (240–279)	482 (5.5%)	92 (1.5%)	24 (0.5%)	35 (1.0%)	633 (2.8%)
Diseases of the circulatory system, ie, ischemic heart diseases, diseases of pulmonary circulation, dysrhythmias, heart failure, cerebrovascular diseases, etc. (390–459)	1,068 (12.3%)	1,380 (23.0%)	3,758 (79.8%)	2,725 (78.8%)	8,931 (39.1%)
Pulmonary diseases, ie, pneumonia and influenza, chronic obstructive pulmonary disease, etc. (460–519)	1,707 (19.6%)	126 (2.1%)	121 (2.6%)	114 (3.3%)	2,068 (9.0%)
Diseases of the digestive system (520–579)	1,302 (15.0%)	756 (12.6%)	75 (1.6%)	96 (2.8%)	2,229 (9.7%)
Diseases of the genitourinary system, ie, nephritis, nephrotic syndrome, nephrosis, and other diseases of the genitourinary system (580–629)	338 (3.9%)	55 (0.9%)	9 (0.2%)	46 (1.3%)	448 (2.0%)
Trauma (800–959)	235 (2.7%)	1,972 (32.8%)	86 (1.8%)	43 (1.2%)	2,336 (10.2%)
Poisoning by drugs and biological substances (960–979)	293 (3.4%)	22 (0.4%)	4 (0.1%)	15 (0.4%)	334 (1.5%)
Other	1,305 (15.0%)	792 (13.2%)	388 (8.2%)	241 (7.0%)	2,726 (11.9%)
Total	8,700 (100.0%)	6,004 (100.0%)	4,707 (100.0%)	3,459 (100.0%)	22,870 (100.0%)

ICD-9, International Classification of Diseases, 9th Revision; MIMIC-II, Multiparameter Intelligent Monitoring in Intensive Care II; MICU, medical intensive care unit; SICU, surgical intensive care unit; CSRU, cardiac surgery recovery unit; CCU, coronary care unit.

Table 4. Incidence and associated mortality of diseases defined by laboratory abnormalities

Type of Pathology	Laboratory Abnormality	No. of ICU Stays With at Least One Valid Measurement[a]	No. of ICU Stays With at Least One Episode of Abnormality	Associated Mortality Rate (Positive Cases)	Associated Mortality Rate (Negative Controls)	Odds Ratio	95% CI
Metabolic acidosis	Arterial pH <7.2 and HCO_3 <18	12,894	785	53.1%	12.7%	7.8	6.7–9.0
Thrombocytopenia	Platelet count <50,000	18,917	872	42.2%	9.8%	6.7	5.8–7.8
Renal insufficiency	Creatinine ≥2 mg/dL	18,855	3,523	25.8%	8.2%	3.9	3.5–4.3
Hypoglycemia	Glucose <40	18,261	163	31.9%	11.6%	3.5	2.5–4.9
Hepatitis	Bilirubin >4.0 OR AST > 500 OR ALT > 500	8,364	1,327	35.8%	14.7%	3.3	2.9–3.7
Leukocytosis	White blood count >15,000	18,509	7,079	18.6%	7.1%	3.0	2.7–3.2
Acute lung injury	Pao_2/Fio_2 ≤300 mm Hg	10,476	6,819	19.8%	12.5%	2.0	1.8–2.3
Anemia	Hematocrit <25	19,046	4,835	15.0%	10.0%	1.6	1.4–1.8
Myocardial infarction	Troponin T >0.1 ng/dL or Troponin I >2.0 ng/dL	5,022	2,904	21.3%	18.9%	1.2	1.0–1.4

ICU, intensive care unit; CI, confidence interval; AST, aspartate aminotransferase; ALT, alanine transaminase.

[a]Excludes hospitalizations with multiple ICU stays and the last 12 hrs of every patient record to avoid end-of-life physiology. Note that odds ratio applies only to ICU stays in which the diagnostic test was performed.

Although MIMIC-II is smaller than Project Impact based on the number of ICU patients, MIMIC-II is notable for its relatively complete ICU patient records and free availability.

DISCUSSION

We have described the development of a large ICU database that is freely accessible to clinical researchers. The highly automated methods of aggregating thousands of ICU records from disparate sources, the use of open-source data formats, and the development of Health Insurance Portability and Accountability Act-compliant distribution mechanisms (with minimal credentialing requirements and no associated fees) are all intended to provide a valuable research resource to the widest audience of users. This database provides a high-resolution record of the dynamics of a patient's pathophysiology and the contemporaneous therapeutic interventions. The interplay between disease and therapy can thus be analyzed. Furthermore, because the data are already electronic, this data-

Table 5. Incidence and associated mortality of abnormalities of individual vital signs

Vital Sign Variable	No. of ICU Stays With at Least One Valid Measurement[a]	Median of Median of Stays (IQR of Median of Stays)	Vital Sign Instability Definition	No. of ICU Stays With at Least One Episode of Instability	Associated Mortality Rate (Positive Cases)	Associated Mortality Rate (Negative Controls)	Odds Ratio	95% CI
Heart rate, beats/min	20,399	84 (74.5–93.5)	HR >150	1,084	32.4%	9.9%	4.4	3.8–5.0
			HR >120	5,820	21.8%	6.8%	3.8	3.5–4.2
			HR <40	294	26.5%	10.8%	3.0	2.3–3.9
Systolic BP, mm Hg	20,398	119 (108–132)	SBP >220	475	20.8%	10.8%	2.2	1.7–2.7
			SBP > 200	1,506	19.0%	10.4%	2.0	1.8–2.3
			SBP <90	10,393	16.9%	5.0%	3.9	3.5–4.3
			SBP <80	5,808	23.8%	6.0%	4.9	4.5–5.4
Respiratory rate, breaths/min	20,347	18 (16–21)	RR >40	2,431	21.3%	9.7%	2.5	2.3–2.8
			RR >30	7,991	18.3%	6.4%	3.3	3.0–3.6
Pulse oximetry	20,383	98 (97–99)	SpO_2 <90%	6,372	21.9%	6.1%	4.3	3.9–4.7
			SpO_2 <80%	2,232	31.2%	8.6%	4.8	4.4–5.4

ICU, intensive care unit; IQR, interquartile range; CI, confidence interval; BP, blood pressure; HR, heart rate; SBP, systolic blood pressure; RR, respiratory rate.

[a]Excludes hospitalizations with multiple ICU stays; the last 12 hrs of every patient record (to avoid end-of-life physiology); and any documented values outside of the following ranges: HR 20–300, SBP 30–250; RR 2–80; SpO_2 30–100.

Table 6. Comparison of MIMIC-II with other ICU databases[a]

	MGH Waveform DB (24)	SIMON (25)	IMPROVE (26)	Project IMPACT (2)	eICU Research Institute (27)	APACHE (1)	ICNARC (28)	Veteran's Administration (29)	MIMIC-I (30)	MIMIC-II
Category of database	OR/cardiac procedures	Trauma ICU	ICU shock and neurologic monitoring)	MICU SICU Trauma CCU CSRU	Adult critical care	Medical and surgical ICU	Medical and surgical ICU	Medical and surgical ICU	MICU SICU CCU	MICU SICU CCU CSRU
Number of records	100+	1,000+	100+	100,000+	1,000,000+	110,558	900,000	110,000+	100+	10,000+
Single- or multicentered	Single	Single	Single	Multiple	Multiple	Multiple	Multiple	Multiple	Single	Single
Source availability	Academic Free	Academic Restricted	Academic Fee-based	Private Fee-based	Private Restricted	Private Free	Academic Free	Government Free	Academic Free	Academic Free
Record length	<90 mins	Entire stay	24 hrs	First 24 hrs	Entire stay	First 7 ICU days	First 24 hrs	Entire stay	Variable (24–48 hrs)	Entire stay
Physiological waveforms	Multichannel: ECG Hemodynamic respiratory	No	Multichannel: ECG Hemodynamic respiratory EEG	No	No	No	No	No	Multichannel: ECG Hemodynamic respiratory	Multichannel: ECG Hemodynamic respiratory
Vital signs and numerics	Monitor-generated (1-sec resolution)	Monitor-generated (1-sec resolution)	Monitor-generated (1-sec resolution)	Nurse daily Abstraction	Monitor-generated (5-min resolution)	Nurse daily Abstraction	Nurse daily Abstraction	Nurse daily Abstraction	Monitor-generated (1-sec resolution)	Monitor-generated (1-sec and minute resolution)
Bedside alarms	Yes	No	Yes	No	No	No	No	No	Yes	Yes
Laboratory/clinical data	No	No	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes
Clinician notes	No	No	Yes	No	No	No	No	Yes	No	Yes
Therapy details	No	No	Yes	Yes	Yes	Daily Abstraction	Daily Abstraction	Detailed Abstraction	No	Yes
ICD-9 Codes/problem lists	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mortality/outcomes data availability	No	No	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes

MGH Waveform DB, Massachusetts General Hospital/Marquette Foundation Waveform database; SIMON, Signal Interpretation and Monitoring; IMPROVE, IMPROVing control of patient status in critical carE; eICU, electronic intensive care unit; APACHE, Acute Physiology and Chronic Health Evaluation; ICNARC, Intensive Care National Audit & Research Centre; MIMIC, Multiparameter Intelligent Monitoring in Intensive Care; ICU, intensive care unit; ICD-9, International Classification of Diseases, 9th Revision; OR, operating room; ECG, electrocardiogram; MICU, medical intensive care unit; SICU, surgical intensive care unit; CCU, coronary care unit; CSRU, cardiac surgery recovery unit.

[a]Category of database describes patient populations of each respective database. Number of records is number of records, based on published or private communication, to the nearest order of magnitude. Single- or multicentered databases are designated by the number of medical centers that contributed records. Ownership that controls access to a database is defined as either academic (university or professional society) or private (corporation). Availability is defined as "free" if potential users are not required to pay access fees and "fee-based" if access fees are required and is considered "restricted" if third-party researchers are generally not granted access to such a database. Record length is defined as the typical length of each ICU record contained in the database. Physiological waveforms describe the types of high-resolution waveform data available with each database, ie, ECG, hemodynamics (eg, pulsatile blood pressure waveforms), electroencephalogram, respiration. Vital signs and numerics recording resolution are described for each database, from second-to-second to daily abstraction (one set of vital signs per day). Clinician notes include one or more of the following: physician progress notes, nursing progress notes, and discharge summaries. Therapy details include daily medication lists, provider order entry records, or medication flow sheets.

base naturally supports the development of clinical decision-support systems, which are automated algorithms that provide alerts, early warnings, and other decision support for critical care. Because the data are available online, along with a comprehensive user guide, it is hoped that an online community of MIMIC-II researchers will develop in which ideas can be exchanged and collaborations can develop.

To our knowledge, MIMIC-II is the only ICU database that encompasses patient demographics, clinical laboratory data, categorical admission diagnoses as well as detailed therapeutic profiles such as intravenous medication drip rates and hourly fluid balance trends for the duration of the ICU stay. These data are supplemented with a rich set of text-based records, including nursing progress notes, discharge summaries, and radiology interpretations. Developing this database was possible only because Boston's Beth Israel Deaconess Medical Center is one of the <5% of hospitals in the United States that have fully automated and comprehensive medical records (10). Furthermore, in the current version of MIMIC-II, physiological waveforms and minute-to-minute vital-sign are available for 2061 of the patient records (an additional 2000 records not matched to clinical data are also available).

A wide range of analyses can be performed on these data, spanning epidemiology, clinical decision-rule development, and electronic tool development. For example, Jia et al (11) assessed risk factors for acute respiratory distress syndrome in MIMIC-II patients who were mechanically ventilated for >48 hrs. Saeed (12) studied how certain ICU practices varied significantly as a function of time of day, ie, during the workday vs. the overnight shifts. From the parameter-rich MIMIC-II database, Hug (13) identified multivariate factors associated with death, successful wean of pressor infusions within 12 hrs, successful weans of intra-aortic balloon pumps, and development of septic hypotension; and he developed predictive statistical models for these outcomes. Because of the suitable size of the database, for example, Hug identified >50,000 episodes of successful pressor weans within MIMIC-II; he was able to segment the data into distinct training and testing subpopulations, which enhances the validity of such analyses.

Broadly speaking, MIMIC-II supports the development of new automated clinical decision-support systems. Although decision-support research before MIMIC-II has spanned functionality, including automated drug allergy notifications, medication interactions, and reminders about abnormal laboratory results (14, 15), little progress has been made in clinical decision support for the acute management of unstable patients, a major challenge in the ICU. The value proposition is that novel automated algorithms may operate in real time and prevent medical mistakes or promote timely responses to the patients' conditions. Consider that Kumar et al (16) showed that the duration of hypotension before the initiation of antimicrobial therapy was the most significant factor associated with mortality in septic patients. Therefore, having a reliable automated decision support to prompt timely antibiotic administration may be expected to improve critical care outcomes.

Distinct components of the MIMIC-II database may be studied to develop a variety of software tools. For example, electronic algorithms that automatically extract information from free-text nursing notes and discharge summaries have been developed and tested (7). Alarm algorithms can be trained and validated using MIMIC-II waveform records, which may help to address the perennial problem of false alarms by bedside monitors (17). Analysis of MIMIC-II also promotes the use of effective mathematical techniques for quantifying patterns through time, because so many of its clinical parameters are complex time-series, eg, continual heart rate trends. Saeed and Mark (18) explored the use of wavelet transformation of hemodynamic time series with machine learning to predict hemodynamic deterioration in ICU patients, and Lehman et al (19) explored how to search for case records who shared similar temporal patterns in time-series variables.

Interpretive algorithms can be developed and compared head-to-head using MIMIC-II. For instance, there are a number of different competing algorithms that estimate cardiac output from the arterial blood pressure waveform. MIMIC-II provides a resource for their fair comparison (20), providing a large number of radial arterial blood pressure waveforms and paired measurements of cardiac output by a thermodilution reference method. MIMIC-II may serve a role anal-

ogous to the public access arrhythmia databases that played an indispensable role in the development, refinement, and—ultimately—widespread acceptance of automated algorithms for electrocardiogram analysis (21). Overall, MIMIC-II offers the means to develop and assess cutting-edge algorithms, exploiting the full spectrum of data available in critical care, with the underlying goal to catalyze a new generation of automated decision-support systems that demonstrably improve the practice of critical care.

In this report, we characterized MIMIC-II using ICD-9 codes as well as quantitative data-driven measures. Although ICD-9 is the accepted coding procedure used for patient billing, prior studies have suggested that ICD-9 administrative data do not accurately reflect the true prevalence of comorbidities in hospitalized patients (22). Some analyses of MIMIC-II may require chart review by clinicians to optimize accuracy. Our group is actively investigating the application of natural language processing technology to automatically identify patients with specific comorbidities such as AIDS, metastatic cancer, chronic obstructive pulmonary disease, etc, that are needed for such scores as SAPS, Acute Physiology and Chronic Health Evaluation, etc.

MIMIC-II was developed to serve as a research resource for physicians, scientists, and engineers. If large volumes of medical data are to be widely and freely disseminated, patient privacy concerns are inevitably raised. To address this, automated deidentification tools were developed and rigorously evaluated to remove protected health information from structured and free-text fields such as nursing notes and physician discharge summaries. Ultimately, we successfully applied tools that were demonstrated to perform better than using two independent clinicians to identify protected health information in medical records. Finally, as an added layer of protection to patient privacy, MIMIC-II users must sign and abide by a data-use agreement before being granted access to the free-text elements to ensure data are used for only legitimate purposes.

Although MIMIC-II is in many respects innovative and unprecedented, it by no means represents the ultimate ICU research database. There are several notable limitations, and ideally future iterations of MIMIC-II or other complimentary public-access ICU databases can

address some of these matters. First, administration of oral and intravenous bolus medication administration was paper-charted and not systematically tabulated in the electronic record. Although provider order entry records were computerized and aggregated into MIMIC-II, these data are not equivalent to documenting if and when those medications were truly administered. Second, MIMIC-II only includes ICU data (with the exception of laboratory results and discharge summaries). A database that included complete data before the development of critical illness would be invaluable. Third, MIMIC-II is limited by having ICU records from a single institution. The clinical practices and patient populations documented by this database may not be representative of other hospitals. Fourth, MIMIC-II data reflects "real-world" clinical practice as opposed to scrupulously tended research protocols. This means that certain documentation and clinical practices may be less reliable (eg, not carefully recalibrating the arterial pressure transducers every shift), which may be a source of error for some analyses, whereas it may be advantageous for other investigations. For example, when developing clinical decision rules and other automated decision-support algorithms, it is more valid to analyze "real-world" data rather than idealized research data that are unrepresentative of actual clinical practice.

Finally, only a subset (approximately 2000) of MIMIC-II records include matched physiological waveform and minute-to-minute trend data owing to technical difficulties in deploying data-archiving machines, which interfaced with bedside patient monitoring systems, and difficulties in linking waveform files to specific clinical records. Furthermore, the subset of records with waveforms is not necessarily statistically similar to the database as a whole. On the other hand, the collection of 2000 waveform records is massive by most standards. For example, the electrocardiogram database that supported development and evaluation of automated arrhythmia algorithms included much smaller data collections (48 half-hour samples in the MIT-BIH Arrhythmia Database [21] and 80 3-hr records in the American Heart Association ECG Database [23]). Machine learning strategies certainly benefit as the number of examples rise, but the 2000 waveform records matched with clinical cases are sufficient for many important investiga-tions, and this number will grow, because we continue to collect data and to publish new versions of MIMIC-II. Also, an additional 2000 records unmatched to clinical data are available (posted on PhysioNet) and such records are adequate for many physiological and clinical studies (such as developing algorithms to predict hypotensive episodes, reducing false alarms, etc).

In the short term, collaboration with industry vendors is mandatory for the development of databases similar to MIMIC-II owing to the sophisticated interfaces and proprietary data formats of most ICU devices. For MIMIC-II, the participating patient monitoring vendor (Philips Healthcare) provided significant engineering resources to facilitate access to data from patient monitors and to the CareVue clinical information systems. In the long term, however, the adoption of common data formats that allow for seamless device communication would remove a significant barrier to developing databases similar to MIMIC-II.

CONCLUSIONS

MIMIC-II documents a diverse and very large population of ICU patient stays and contains comprehensive and detailed clinical data, including physiological waveforms and minute-by-minute trends for a subset of records. It establishes a new public-access resource for critical care research, supporting a diverse range of analytic studies spanning epidemiology, clinical decision-rule development, and electronic tool development.

REFERENCES

1. Zimmerman JE, Kramer AA: Outcome prediction in critical care: The Acute Physiology and Chronic Health Evaluation models. *Curr Opin Crit Care* 2008; 14:491–497
2. Higgins TL, Teres D, Nathanson B: Outcome prediction in critical care: The mortality probability models. *Curr Opin Crit Care* 2008; 14:498–505
3. The Acute Respiratory Distress Syndrome Network: Ventilation with lower tidal volumes as compared with traditional tidal volumes for acute lung injury and the acute respiratory distress syndrome. *N Engl J Med* 2000; 342:1301–1308
4. Mehta R, Kellum J, Shah S, et al: Acute Kidney Injury Network: Report of an initiative to improve outcomes in acute kidney injury. *Crit Care* 2007; 11:R31
5. Dellinger R, Levy M, Carlet J, et al: Surviving Sepsis Campaign: International guidelines for management of severe sepsis and septic shock: 2008. *Intensive Care Med* 2008; 34: 17–60
6. The WFDB Software Package. Available at: http://www.physionet.org/physiotools/wfdb.shtml. Accessed September 6, 2010
7. Neamatullah I, Douglass MM, Lehman LW, et al: Automated de-identification of free-text medical records. *BMC Med Inform Decis Mak* 2008; 8:32
8. De-identification: Software and test data. Available at: http://www.physionet.org/physiotools/deid/. Accessed September 6, 2010
9. MIMIC-II. Version 2.4. Available at: http://mimic.physionet.org/database/releases/version-24.html. Accessed September 6, 2010
10. Jha AK, DesRoches CM, Campbell EG, et al: Use of electronic health records in US hospitals. *N Engl J Med* 2009; 360:1628–1638
11. Jia X, Malhotra A, Saeed M, et al: Risk factors for ARDS in patients receiving mechanical ventilation for >48 h. *Chest* 2008; 133: 853–861
12. Saeed M: Temporal pattern recognition in multiparameter ICU data. PhD thesis, Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, 2007. Available at: http://dspace.mit.edu/handle/1721.1/40507. Accessed January 20, 2011
13. Hug CW: Predicting the risk and trajectory of intensive care patients using survival models. PhD thesis, Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, 2009. Available at: http://dspace.mit.edu/handle/1721.1/33957. Accessed January 20, 2011
14. Wright A, Sittig DF, Ash JS, et al: Clinical decision support capabilities of commercially-available clinical information systems. *J Am Med Inform Assoc* 2009; 16:637–644
15. Gardner R, Shabot M: Computerized ICU data management: Pitfalls and promises. *Int J Clin Monit Comput* 1991; 7:99–105
16. Kumar A, Roberts D, Wood K, et al: Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. *Crit Care Med* 2006; 34:1589–1596
17. Aboukhalil A, Nielsen L, Saeed M, et al: Reducing false alarm rates for critical arrhythmias using the arterial blood pressure waveform. *J Biomed Inform* 2008; 41: 442–451
18. Saeed M, Mark R: A novel method for the efficient retrieval of similar multiparameter physiologic time series using wavelet-based symbolic representations. *AMIA Annu Symp Proc* 2006:679–683
19. Lehman L, Saeed M, Moody G, et al: Similarity-based searching in multi-parameter time series databases. *Comput Cardiol* 2008; 35:653–656
20. Sun JX, Reisner AT, Saeed M, et al: The cardiac output from blood pressure algorithms trial. *Crit Care Med* 2009; 37:72–80

21. Moody G, Mark R: The impact of the MIT-BIH arrhythmia database. *IEEE Eng Med Biol Mag* 2001; 20:45–50

22. Quan H, Parsons GA, Ghali WA: Validity of information on comorbidity derived from ICD-9—Critical care medicine administrative data. *Med Care* 2002; 48: 675–685

23. American Heart Association ECG Database DVD. Available at: http://www.ecri.org/Products/Pages/AHA_ECG_DVD.aspx. Accessed September 6, 2010

24. The MGH/MF waveform database. Available at: http://www.physionet.org/pn3/mghdb/. Accessed September 6, 2010

25. Norris P, Dawant B: Closing the loop in ICU decision support: Physiologic event detection, alerts, and documentation. *Proc AMIA Symp* 2001:498–502

26. Nieminen K, Langford R, Morgan C, et al: A clinical description of the IMPROVE data library. *IEEE Eng Med Biol Mag* 1997; 16:21–24

27. Celi L, Hassan E, Marquardt C, et al: The eICU: It's not just telemedicine. *Crit Care Med* 2001; 29:183–189

28. Harrison DA, Parry GJ, Carpenter JR, et al: A new risk prediction model for critical care: The Intensive Care National Audit & Research Centre (ICNARC) model. *Crit Care Med* 2007; 35:1091–1098

29. Render ML, Deddens J, Freybert R, et al: Veterans Affairs intensive care unit risk adjustment model: Validation, updating, recalibration. *Crit Care Med* 2008; 36: 1031–1042

30. Moody G, Mark R: A database to support development and evaluation of intelligent intensive care monitoring. *Comput Cardiol* 1996; 33:657–660