Appendix C

The Evaluation Data Set

This appendix contains documents that describe the evaluation experiment. Section C.1 contains the original call for participation that was circulated to solicit volunteer subjects for the experiment. Section C.2 contains the instructions that were provided to the volunteer subjects before they began the database-design task. Section C.3 contains the complete clinical data set which the database-design teams were asked to model in a relational database schema.

C.1 Call for Participation in Experiment

The following call for participation was circulated via electronic mail to the student bodies of the medical information sciences department at Stanford Medical School and the database research group at the Stanford Computer Science Department. The database-design teams for the experiment were composed of the respondents to this message.

> Greetings,
>
> I will be conducting an experiment as part of my thesis evaluation and need a few good women and men. You are qualified to participate if you have clinical experience and/or experience in relational database design.
>
> The experiment will involve several teams of two members--one knowledgeable in the medical domain, the other in database design--developing and implementing relational database schemas to represent a clinical data set that I will provide. The data set represents

clinical data for a single patient over a two day period and is of modest scope (roughly 40 data elements) and of modest size (roughly 150 recorded values). Hence, the database design task is much smaller than that of a full-blown clinical database, though it is representative of the modeling challenges inherent in the design of a real database. Each team will design a database schema independently to represent the clinical data set, and subsequently I will apply my abstract data model and query language to map across the different databases. Each team will also encode a small set of questions that I will provide into SQL queries that can execute against its database.

I've included an abstract describing my work and its objectives at the end of this message, and I would be glad to talk to anyone who has more questions. Your reward for participating in the experiment, in addition to my everlasting gratitude and an all-expense-paid trip to Tony and Alba's pizzeria, will be the opportunity to design your own clinical database schema with a partner whose expertise in this task complements your own. MIS students with a clinical background who are considering taking CS245 this winter (Database System Implementation) may be especially interested, as this course includes a database project.

If you decide to participate, you will specifically be asked to do the following:

1. Based on a clinical data set that I provide, design a relational database schema to capture the information in the data set and to support SQL queries against the data. The data that I provide will be in a very general format and you will need to make a number of modeling decisions in designing the relational schema.

2. Implement the schema you have designed in Sybase DBMS and populate the database with the clinical data that I have provided.

3. Translate a modest number (5-10) of informal questions that I provide into formal SQL queries that will answer the questions based on your relational database. For example, a question may be "what were all of the temperature readings for the patient named Chris Webber between 12:00 and 5:00 PM." The questions will be relatively simple.

Participating in the experiment will require approximately 10-15 hours of your time over a 3 month period beginning in November. Most of your help will be needed in January, so the experiment will not conflict with final exams and the holidays. Please contact me as soon as possible if you are interested in participating. Thank you.

ABSTRACT:

Bridging the Representational Heterogeneity of Clinical Databases

The goal of our work is to overcome the heterogeneity of clinical databases so as to provide medical decision-support applications with a uniform interface to patient data. Most decision-support applications require patient-specific data because they provide advice in the context of clinical cases. These applications include alerting systems, such as the HELP system and the Regenstrief Medical Record System; diagnostic systems, such as MYCIN and QMR; and information-retrieval systems, such as MEDLINE. The emergence and expansion of on-line clinical databases provides the potential for applications to access patient-specific data automatically and thereby to streamline the decision-support process by relieving users of the data-entry burden. The benefits of this mode of interaction have been demonstrated in existing systems [Kuperman, 1991, McDonald, 1984]. However, applications that automatically access patient data currently cannot be shared among health-care institutions because the means by which clinical data are represented and retrieved vary widely. For example, the help system's decision-support software in operation at the Latter-day Saints Hospital cannot be ported to other provider sites without significant modifications because the database queries generated by the help system are specific to the database structure, the query language, and the coding scheme of the help database. Disparities in the representation and retrieval of data across clinical databases exist, although the information stored at different sites is conceptually quite similar. The diverse implementation of databases containing conceptually similar information is termed representational heterogeneity. Representational heterogeneity requires that database queries be recoded manually when decision-support applications are ported from one provider site to another. The recoding of queries is necessary even for applications that use ostensibly portable knowledge-representation languages, such as the Arden Syntax [Hripcsak, 1990], because existing languages provide no standard model of clinical data. Manual recoding of queries not only entails significant effort, but also introduces a source of error because the semantics of queries may not be preserved when the queries are reformulated in a different environment.

We are developing a methodology, called TransFER, to bridge the representational heterogeneity of clinical databases and to obviate the need for manual recoding of database queries. The general strategy of TransFER is to define a global and uniform reference model of clinical information against which applications' requests for patient data may be formulated. The reference model is abstract, and is independent of any clinical database

implementation. A formal mapping language allows database administrators to specify the correspondence between the clinical database implementation at their sites and the abstract reference model. Currently, only clinical databases implemented using the relational model can be accommodated by the mapping language. A translating compiler at each site uses the encoded mappings to translate automatically queries specified against the global model to queries that can be executed locally and that are semantically equivalent.

We propose to use the ICU data set to evaluate the generality and utility of the TransFER query translation methodology. Our research hypothesis is that the TransFER methodology can map the reference model effectively to three independently designed relational database schemata, all of which represent the same data set, and that the TransFER methodology can translate automatically queries formulated against the single reference model to queries that can be executed against all three schemata. The experiment will allow us to evaluate the TransFER methodology in the context of a realistic clinical data set and with respect to independently implemented databases.

References:

Hripcsak, G., Clayton, P. D., Pryor, T. A., Haug, P., Wigertz, O. B.& Van der lei, J. (1990). The Arden Syntax for Medical Logic Modules. In R. Miller (Ed.), Proceedings of the Fourteenth Symposium on Computer Applications in Medical Care . Washington, D.C., pp. 200-204.

Kuperman, G. J., Gardner, R. M.& Pryor, T. A. (1991). HELP: A Dynamic Hospital Information System. New York: Springer-Verlag.

McDonald, C. J. (1984). Reminders to Physicians from an Introspective Computer Medical Record. Annals of Internal Medicine, 100: 130-138.

C.2 Instructions to Database-design Teams

The following document contains the instructions that I provided to the volunteer subjects regarding the database-design task. Three database-design teams participated in the experiment; two of the teams consisted of one computer-science department student and one physician, and one of the teams consisted of a single physician with considerable experience in relational-database design.

The Clinical Database Design Derby

I. General instructions

The task is to represent in a relational database the clinical and demographic information for two inpatients. The data will be provided to you in a form resembling a patient chart. The objective is to generate a relational database schema that can capture the information in the "chart" and to populate the database with this information. The Sybase DBMS will be used for the implementation. The products of the design process include:

1. A set of SQL data definition language commands that define the relational database schema. These command will be specific to Sybase's "flavor" of SQL.

2. Documentation of the codes that are used in the populated database to represent the patient data (this may be part of the database or may be external to it)

3. An implemented and populated Sybase database that can be queried using SQL

II. Designing the database

• You may use any database design technique you wish to define the relational schema (ER diagrams, object-oriented analysis, etc.).

• The final relational schema should be represented as a set of commands in the Sybase data definition language. The commands are similar to the data definition language (DDL) commands of SQL. If you need to familiarize yourself with the Sybase DDL, Sybase documentation is available at the Section on Medical Informatics and elsewhere.

• You may use any names, codes, etc. you wish to represent the table and field names of the schema, consistent with the constraints of Sybase.

• Each table must have a primary key defined, consisting of one or more fields of that table. All foreign keys also must be explicitly defined. The Sybase DDL includes constructs to do this.

III. Populating the database

• After you have finished the database design, the laboratory and flowsheet data will be provided in electronic form. You may either write a small program to load the data into your database, write a program to generate the SQL commands to load the data, or manually write the SQL commands to load the data (or interactively load the data! -- not recommended).

• The history and physical, demographics, and medication list are not available in electronic form. However, this set of data is small.

• You may use any names, codes, etc. you wish to represent the values appearing in the database (except for numerical values, which must appear in their original integer or real form). These codes may be taken from standard terminologies, such as ICD9, CPT4, and SNOMED, or they may be of your own invention. For each name or code whose meaning is not obvious, you must specify either the standard vocabulary from which the code was taken or a more descriptive textual definition of the code.

IV. Rules of the game

1. The final populated relational schema must represent all of the data provided (i.e., *at least* all of the data — it may represent more data as specified below).

2. You may *enrich* the data set by adding types of objects that are not currently represented, as well as instances of those types. However, you may not *expand* the data set by adding instances of any object types that already are represented. For example, you could add data about each patient's insurance coverage because this type of information is not currently represented, but you could not add any chemistry lab results to the set of results already in the data. When in doubt, ask yourself "would adding these data change the results of any valid query issued against the original data set?" If the answer is yes, don't add them.

3. The objective is for the schemas to be designed independently. Therefore, you may not consult with any members of the other database design groups or with any of the physicians who will be formulating queries against the databases. The members of these groups include:

<list of participants omitted>

V. The Data

The following documents will be provided to you.

1. Patient descriptions (demographic and encounter data)

2. Physician descriptions (demographic data)

3. Lab results (the results of chemistry and hematology lab tests)

4. ICU flowsheet data in graphical form (charted physiological parameters captured in the ICU)

If you have any questions at all about the data or about the database-design task in general, please contact me.

Thanks for your participation!

C.3 The Clinical Data Set

The clinical data set consists of demographic, laboratory, medication, and ICU-monitoring data for two patients over the course of approximately 24 hours in the hospital. The data are subdivided into a demographic and history section, a laboratory-test result section, and a charted flowsheet section. In addition to the patient data, there are also physician data that describe the patients' caregivers. The data for Patient #1 were provided by Isaac Kohane and Serdar Uckun as a test-data set to all participants in the 1993 AAAI Artificial Intelligence in Medicine conference.

C.3.1 Patient #1: Jessica Hepburn

Name: Jessica Hepburn
1229 Industrial Ave
Fremont, CA 94520
(415) 320-9934
Medical Record #: 238-94-4227
Attending: Dr. Bill Fold
Admit Date: 8/16/93 17:00

History and Physical

8.5 month old **asian female (date of birth: 12/28/92)** with **biliary atresia,** S/P (surgical procedure) **Kasai procedure on 5/22/93**, **ascending cholangitis** with a history of multiple liver abscesses and portal hypertension.

She was transferred to The Children's Hospital after a **15 day history of E. Coli sepsis**, worsening despite antibiotic therapy. She **currently is in liver failure with coagulopathy**. She has required **mechanical ventilatory support** due to the **ARDS** that followed the onset of sepsis.

Her **admission body weight is 5 kg**.

In addition to medications listed in the ICU flowsheet, the patient receives:

Oral nystatin
Ativan 0.5 mg IV every 6 hours
Amphotericin bladder irrigation every 12 hours
Nebulized ventolin PRN
Maalox 10cc by nasogastric tube every 6 hours for gastric pH < 5

LABS
08/16/93   18:13

Prothrombin Time 19.9 sec   Prothrombin Time - Control 12.0 sec
PTT 68.6 sec   PTT - Control 20 sec
Plasma Fibrinogen 100 mg/dl

RBC 3.78 million/mm^3
MCV 91.9 fl
WBC 11.77 /mm^3
HCT 34.8 %
Mean Corpuscular Hemoglobin Conc 33.8 g/dl
Hemoglobin Conc 11.7 g/dl
Mean Corpuscular Hemoglobin 31.1 pg
Platelet Count 43 /mm^3

Blood Glucose 100 mg/dl
Plasma CO2 21 mEq/L
Serum CL^- 97 mEq/L
BUN 72 mg/dl
Serum K^+ 4.2 mEq/L
Serum Creatinine 1.1 mEq/L
Serum NA^+ 136 mEq/L

LABS
08/16/93   23:16
Serum Ionized Ca^{++} 0.78 mEq/L
PaO2 87 mmHg
PaCO2 44 mmHg
Arterial pH 7.28

LABS
08/17/93   04:30

Arterial O2 Saturation 96 %
Arterial Bicarb Conc 22 mEq/L
Arterial pH 7.23
PaCO2 53 mmHg
PaO2 93 mmHg
Serum Ionized Ca++ 0.90 mEq/L

LABS
08/17/93 04:31
Prothrombin Time 21.8 sec Prothrombin Time - Control 12.0 sec
PTT 76.6 sec PTT - Control 20 sec
Plasma Fibrinogen 89 mg/dl

WBC 13.72 /mm^3, RBC 3.80 million/mm^3
MCV 88.9 fl
Mean Corpuscular Hemoglobin Conc 33.4 g/dl
Hemoglobin Conc 11.3 g/dl
Mean Corpuscular Hemoglobin 29.7 pg
HCT 33.8

SGPT 33 U/L
SGOT 122 U/L
BUN 73 mg/dl
Blood Glucose 97 mg/dl
Plasma CO2 Content 20 mEq/L
Plasma Fibrinogen 89 mg/dl
Serum Albumin 3.4 g/dl
Serum Ca^{++} 9.2 mEq/L
Serum Cl^- 99 mEq/L
Serum Creatinine 1.0 mg/dl
Serum Inorganic Phosphate 4.9 mg/dl
Serum K^+ 4.0 mEq/L
Serum Mg^+ 2.4 mEq/L
Serum Na^+ 137 mEq/L
Serum Total Protein 6.8 g/dl
Serum Triglycerdies 150 mg/dl
Total Bilirubin 29.8 mg/dl
Whole Blood Ammonia 382 mcg/dl

LABS
08/17/93 06:11
Arterial O2 Saturation 95 %
Arterial Bicarb Conc 21 mEq/L
Arterial pH 7.27
PaCO2 45 mmHg
PaO2 87 mmHg

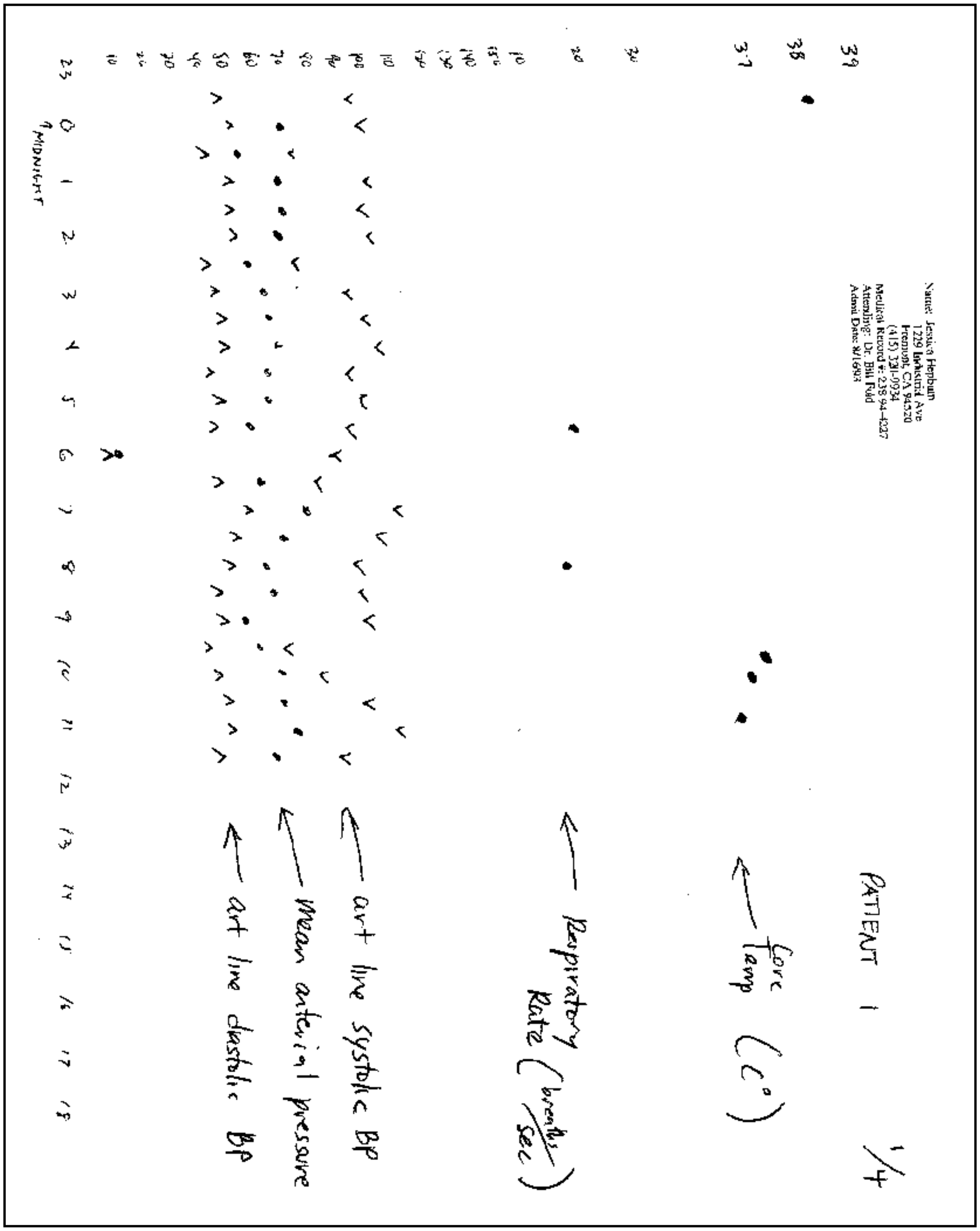
LABS
08/17/93 09:52
Arterial O2 Saturation 97 %
Arterial Bicarb Conc 20 mEq/L
Arterial pH 7.31
PaCO2 41 mmHg
PaO2 95 mmHg
Serum Ionized Ca^{++} 0.82 mEq/L

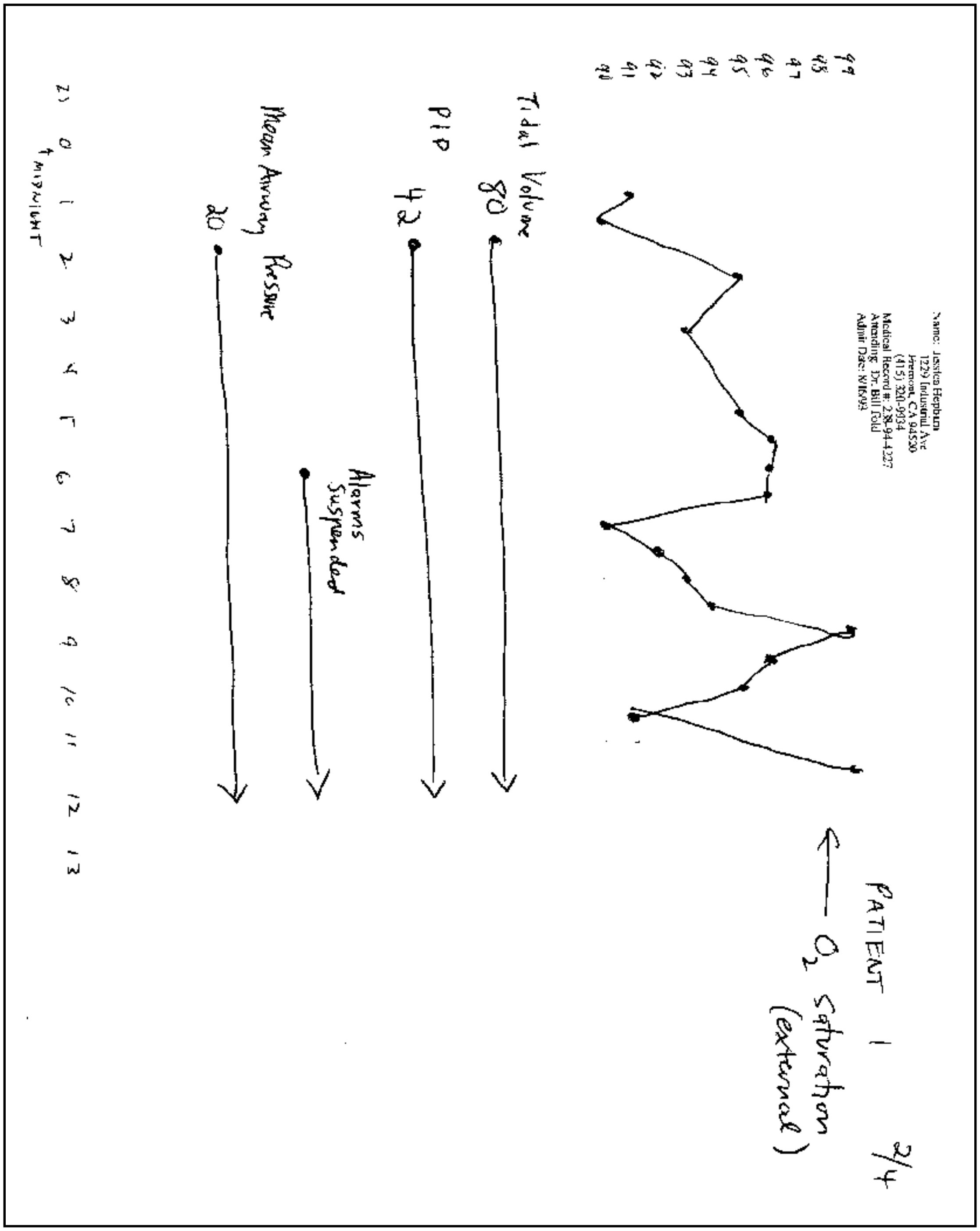
LABS
08/17/93 12:25
Arterial O2 Saturation 94 %
Arterial Bicarb Conc 21 mEq/L
Arterial pH 7.28
PaCO2 44 mmHg
PaO2 80 mmHg
Serum Ionized Ca^{++} 0.87 mEq/L

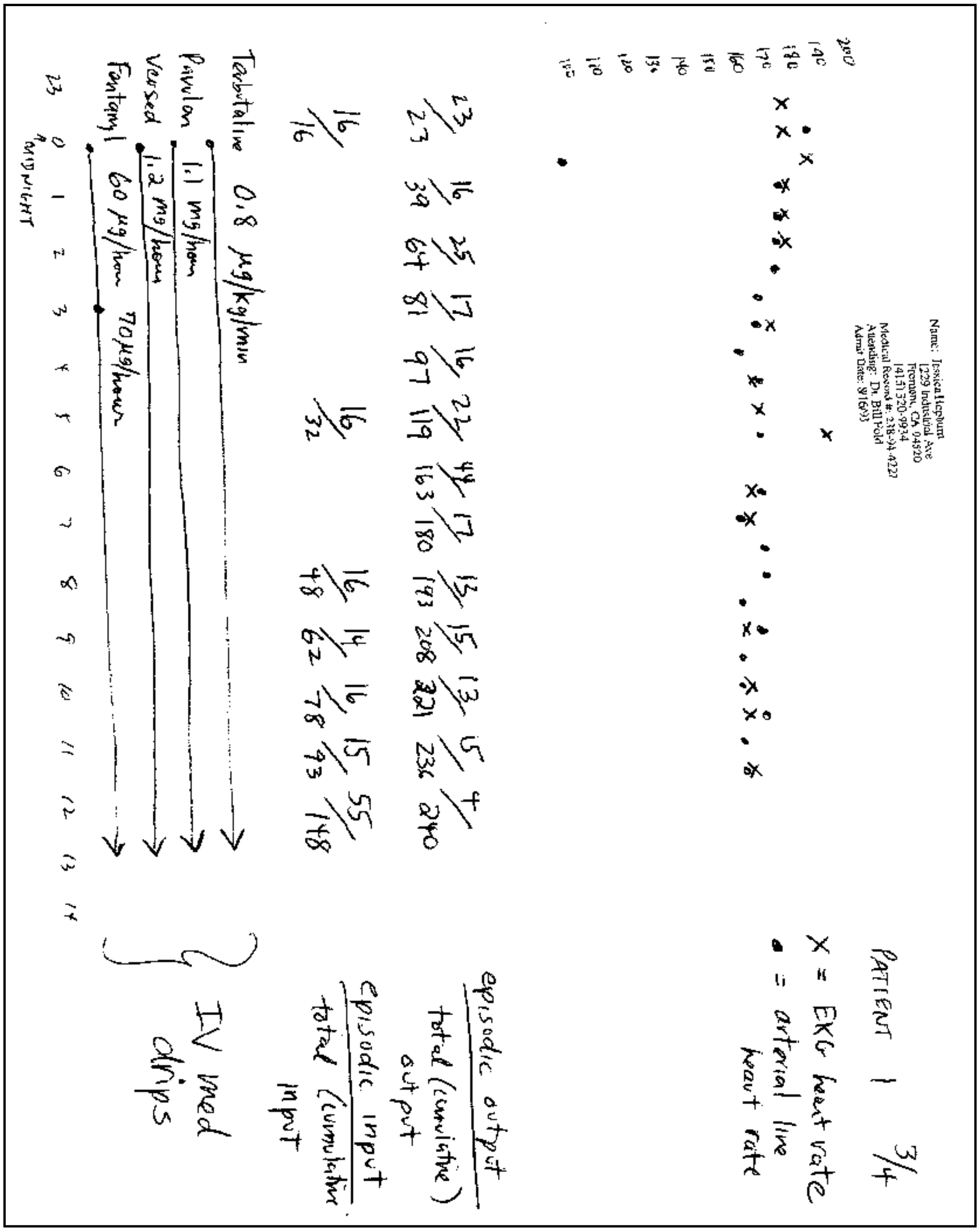
LABS
08/17/93 12:41

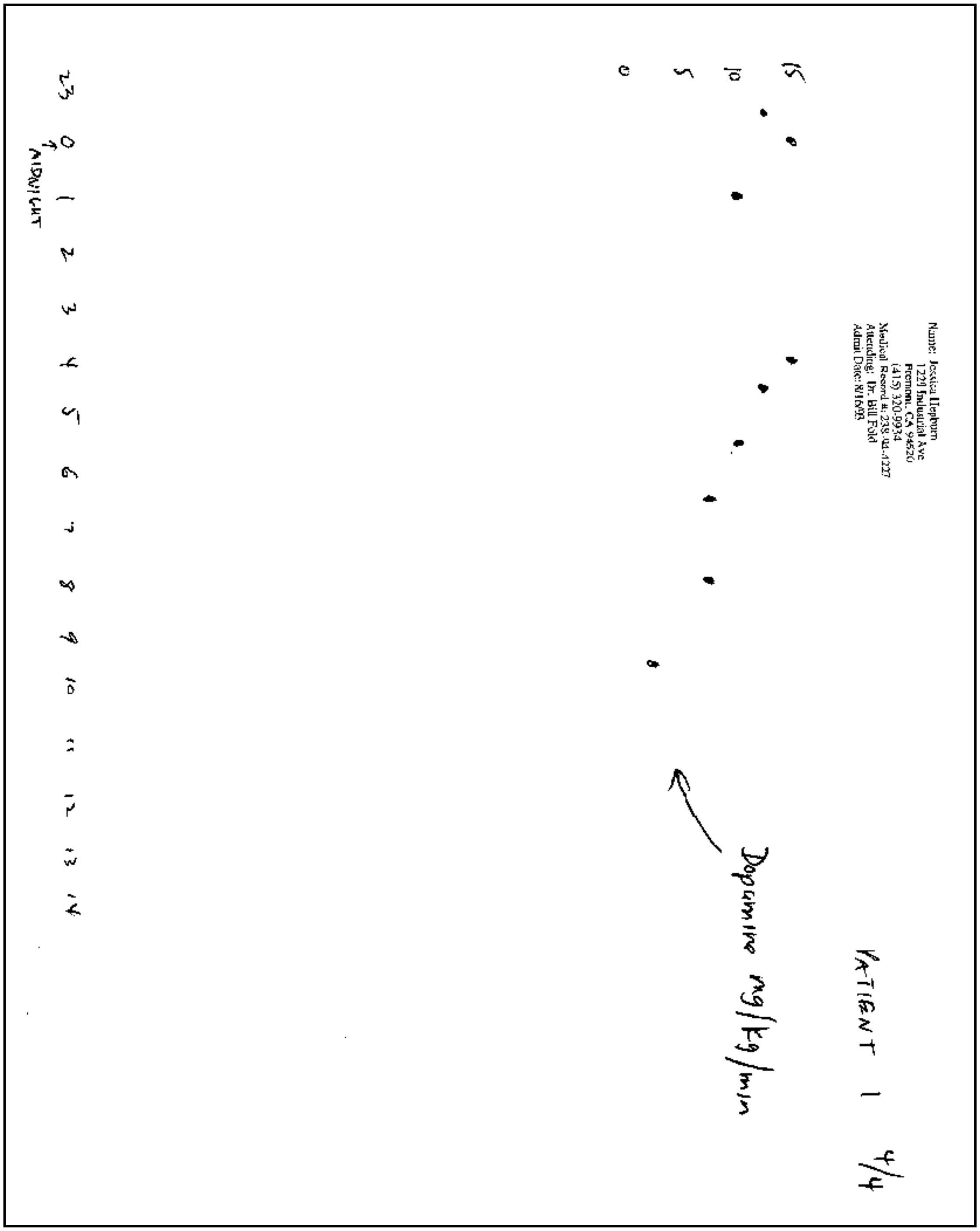
Platelet Count 108 /mm^3

The following four pages contain flowsheet data captured during the patient's stay in the ICU. This data was provided to the database-design teams in exactly the form shown.


PATIENT 1
1/4
Name: Jessica Hepburn
Core Temp (C°)
Respiratory Rate (breaths/sec)
art line systolic BP
Mean arterial pressure
art line diastolic BP
MIDNIGHT


PATIENT 1
2/4
O2 Saturation (external)
Tidal Volume
80
PIP
42
Mean Airway Pressure
20
Alarms Suspended
MIDNIGHT


Name: Jessica Hepburn
1229 Industrial Ave
Fremont, CA 94520
(415) 320-9934
Medical Record #: 238-94-4227
Attending: Dr. Bill Fold
Admit Date: 8/16/93
PATIENT 1 3/4
X = EKG heart rate
• = arterial line heart rate
200
190
180
170
160
150
140
130
120
110
100
23/23 16/39 25/64 17/81 16/97 22/119 44/163 17/180 13/193 15/208 13/221 15/236 4/240
16/16 16/32 16/48 14/62 16/78 15/93 55/148
episodic output / total (cumulative) output
episodic input / total (cumulative input
Terbutaline 0.8 μg/kg/min
Pavulon 1.1 mg/hour
Versed 1.2 mg/hour
Fentanyl 60 μg/hour 70 μg/hour
IV med drips
23 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14
MIDNIGHT


PATIENT 1 4/4
Dopamine mg/Kg/min
MIDNIGHT
15
10
5
0
23 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14
Name: Jessica Hepburn
1229 Industrial Ave
(415) 320-9934
Attending: Dr. Bill Fold

C.3.2 Patient #2: Anna Robik

Name: Anna Robik
415 Treeline St.
Los Altos, CA 94316
(415) 434-6733
Medical Record #: 145-78-3966
Attending: Seth O'Scope, MD
Admit Date: 5/24/93 06:00

History and Physical

Mrs. Robik is a 57 year old **caucasian woman** (**date of birth: 2/14/36**) admitted on 5/24/93 **for exacerbation of asthma**. At that time, her **respiratory rate was 36**, **pulse was 131** and **blood pressure was 150/100**.

Past history includes: **asthma since age 8**, **left nasal polyp removal 1985**, **chronic sinusitis since 1980**, **chronic renal insufficiency since 1988**, and **right renal cyst removal in 1993**.

She had a **normal renal ultrasound on 4/10/93**.

Medications on admission to hospital:

Albuterol MDI 4 puffs q 4
Theodur 100 mg po BID
Azmacort MDI 4 puffs q 4
Pulmoaid nebulizer prn .2 cc Albuterol /.2 cc NS
Brokosol prn
Cholorpheniramine prn

Allergies: Aspirin, non-steroidal antiinflammatory medications

Transfer Note

Her worsening **respiratory distress** led to a **transfer to the ICU on 5/25/93 10:00** where she remains.

ICU MEDS:

D5 NORMAL SALINE @ 100 CC/HOUR
Ranitidine 6 mg/min IV drip
Solumedrol 120 mg IV q 6 hours
Cefotetan 1 gm IV q 12 hours

Admission LABS
5/24/93 06:00 :

Serum NA^+ = 131 mEq/L, Serum K^+ = 4 mEq/L, Serum CL^-=104 mEq/L, Plasma CO2 Content = 28 mEq/L, Serum Creatinine = 1.7 mEq/L, Blood Glucose = 131 mg/dl

WBC=9.6 $/mm^3$, HCT = 45.8 %, Hemoglobin Conc 15.3 g/dl
39% Polymorphonuclear cells, 2% Band Cells, 34% Lymphocyes, 16% Monocytes, 8% Eosinophils

ABG ON 100% FACE MASK: Arterial pH = 7.26, PaCO2 = 56 mmHg,
PaO2 =232 mmHg

Notes in ICU

LABS
5/25/93 13:00
HCT =39 %, WBC = 28 $/mm^3$, Theophylline level = 1.8 µg/ml

ABG ON 100% FACE MASK: Arterial pH = 7.39, PaCO2 = 35 mmHg,
PaO2 =80 mmHg

LABS
5/25/93 15:00
Serum NA^+ = 142 mEq/L, Serum K^+ = 4.1 mEq/L, Serum CL^-=107 mEq/L,
Plasma CO2 Content = 22 mEq/L

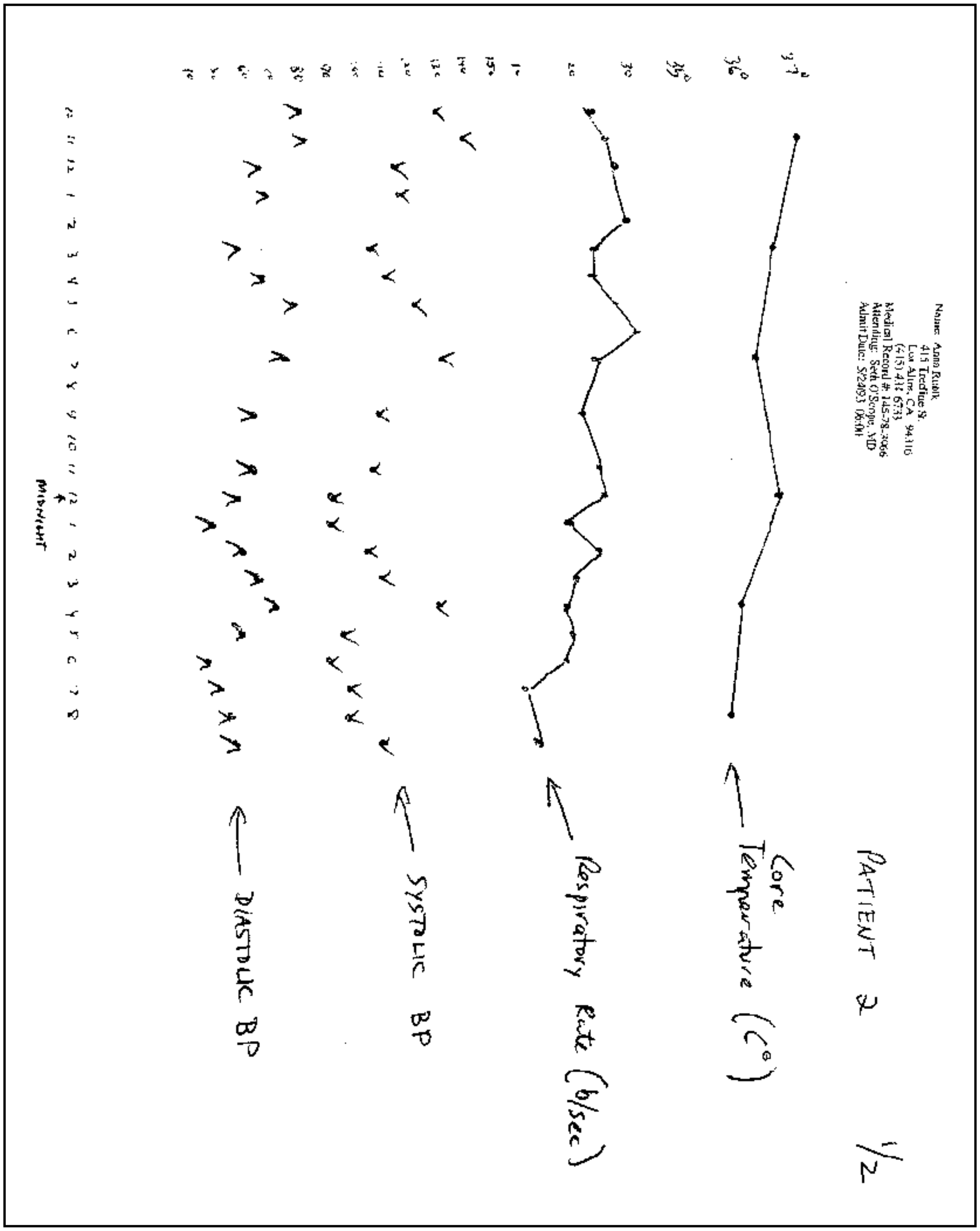
WBC = 19 $/mm^3$ Platelet Count = 138 $/mm^3$, Theophylline level = 1.8 µg/ml

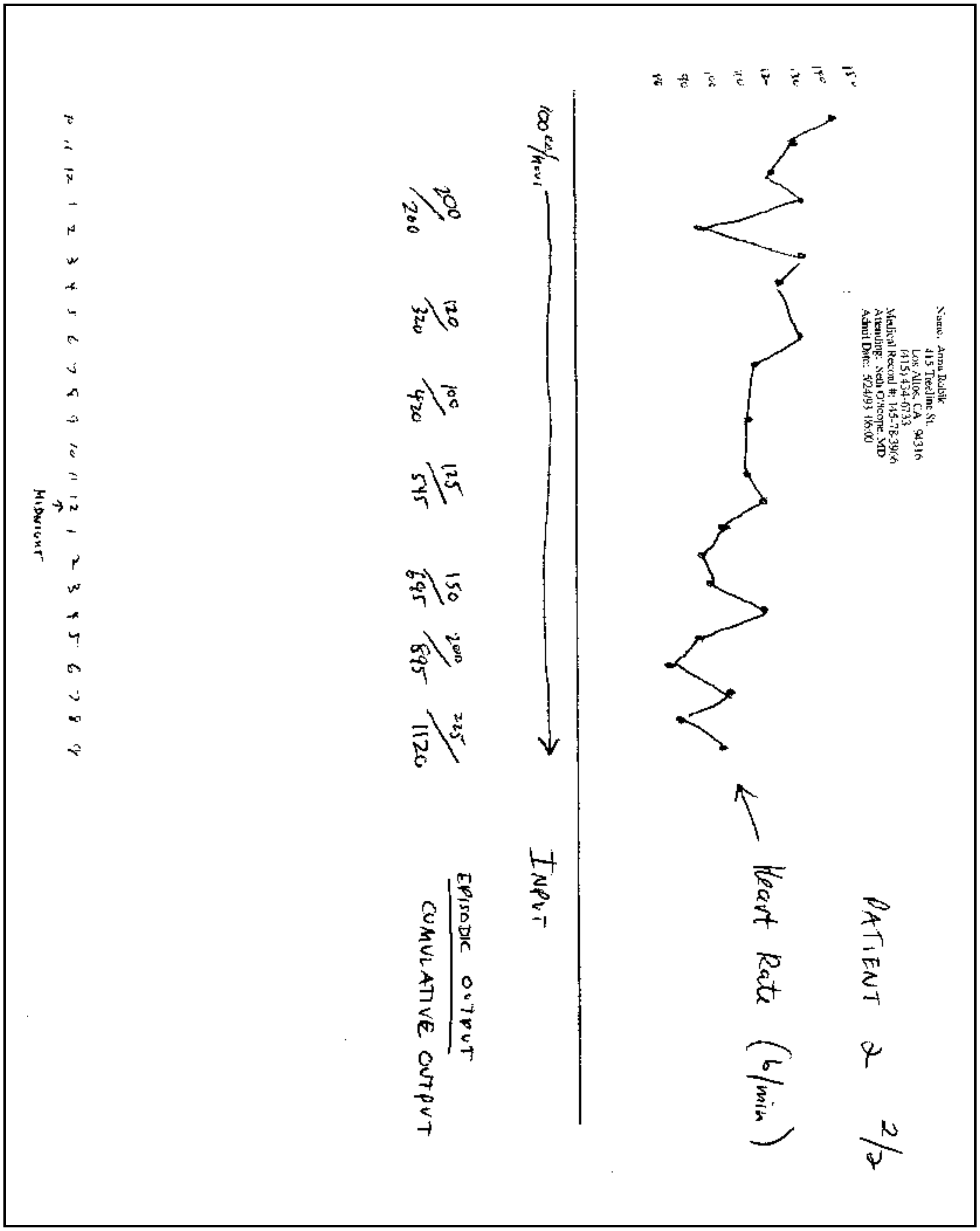
LABS
5/26/93 04:00
Serum NA^+ = 143 mEq/L, Serum K^+ = 4.0 mEq/L, Serum CL^-=103 mEq/L,
Plasma CO2 Content = 23 mEq/L, BUN = 19 mg/dl, Serum Creatinine = 1.3 mEq/L,
Blood Glucose = 147 mg/dl

WBC = 28.5 $/mm^3$, HCT = 37.8 %

Theophylline level = 5.7 µg/ml

The following two pages contain flowsheet data captured during the patient's stay in the ICU. This data was provided to the database-design teams in exactly the form shown.


PATIENT 2
1/2
Core Temperature (C°)
Respiratory Rate (b/sec)
SYSTOLIC BP
DIASTOLIC BP
MIDNIGHT


PATIENT 2 2/2
Heart Rate (b/min)
INPUT
EPISODIC OUTPUT
CUMULATIVE OUTPUT
MIDNIGHT

C.3.2 Physician data

Name: Bill Fold
Date of birth: 5/17/43

Office: 1212 Groupractice Way
Palo Alto, CA 94301
(415) 325-5322
Home: 34 Elm Drive
Los Altos, CA 94233
(415) 495-2144
DEA#: 478832
Specialty: Gastroenterology

Name: Seth O'Scope
Date of birth: 1/26/50

Office: 444 Tallbuilding St.
San Mateo, CA 95300
(415) 744-9003
Home: 12 Waterview Lane
Foster City, CA 94522
(415) 565-3800
DEA#: 597721
Specialty: Internal Medicine