Two open access, high-quality datasets from anesthetic records

David Cumin,[1] Vanessa Newton-Wade,[2] Michael J Harrison,[3] Alan F Merry[4,5]

[1]Centre for Medical and Health Science Education, University of Auckland, Auckland, New Zealand
[2]Information Services, University of Auckland, Auckland, New Zealand
[3]Department of Anaesthesiology, University of Auckland, Auckland, New Zealand
[4]School of Medicine, University of Auckland, Auckland, New Zealand
[5]Department of Anaesthesia, Auckland City Hospital, Auckland, New Zealand

Correspondence to
Dr David Cumin, Centre for Medical and Health Science Education, University of Auckland, Private Bag 92019, Auckland 1142, New Zealand; d.cumin@auckland.ac.nz

Received 8 May 2012
Accepted 21 July 2012
Published Online First 4 August 2012

ABSTRACT

Objective To provide a set of high-quality time-series physiologic and event data from anesthetic cases formatted in an easy-to-use structure.

Materials and methods With ethics committee approval, data from surgical operations under general anesthesia were collected, including physiologic data, drug administrations, events, and clinicians' comments. These data were de-identified, formatted in a combined CSV/XML structure and made publicly available.

Results Two separate datasets were collected containing physiologic time-series data and time-stamped events for 34 patients. For 20 patients, the data included 400 physiologic signals collected over 20 h, 274 events, and 597 drug administrations. For 14 patients, the data included 23 physiologic signals collected over 69 h, with 286 time stamped comments.

Discussion Data reuse potentially saves significant time and financial costs. However, there are few high-quality repositories for accessible physiologic data and clinical interventions from surgical cases. De-identifying records assists with overcoming problems of privacy and storing the data in a format which is easily manipulated with computing resources facilitates access by the wider research community. It is hoped that additional high-quality data will be added. Future work includes developing tools to explore and visualize the data more efficiently, and establishing quality control measures.

Conclusion An approach to collecting and storing high-quality datasets from surgical operations under anesthesia such that they can be easily accessed by others for use in research has been demonstrated.

INTRODUCTION

The advent and proliferation of electronic anesthetic records provides a potential source of data for improving clinical knowledge and practice.[1–4] Much information in electronic anesthetic records could be used in research to improve patient care.[5] Indeed, data from anesthetic records have been used to good effect for research purposes over the years.[3 6–11] A set of generic data would be useful—for example, freely available intraoperative time series of interventional and physiologic data could be used to test diagnostic or alarm algorithms, to estimate sample sizes for certain prospective interventional studies, or to set standards for physiologic and pharmacologic models in human patient simulators.[12] There may be many other applications for such data, particularly if databases were large and their quality high. So far, we know of few publicly accessible databases of this type.

Understandably, the traditional approach to collected data has often been protectionist. Some researchers have been reluctant to share their data as they fear the loss of their competitive edge and intellectual property. Indeed, not all data are suitable for open access. However, we think that some data sources that are difficult to access should be made more accessible. The raw data obtained in research can often be analyzed in ways and for purposes not envisaged by those who collected it. The UK Colonial Registers and Royal Navy Logbooks project provides an excellent example: data from ships logs originally created for navigational purposes and safety are now being used to help predict climate change.[13]

Awareness of the value of open data is growing and there have recently been moves to make it compulsory for publicly funded researchers to share the results and also the data from their research. The NIH has a policy on this,[14] and other publicly funded institutions are following suit: any research funded above a certain sum must have a data management plan and either the data must be made freely available or a compelling reason to the contrary must be provided.[14]

An open approach to data is not without its challenges. Collection of high-quality medical intervention and physiologic data is costly and time consuming. The quality of data varies considerably. For example, the detail of how physiologic signals have been collected (including issues of calibration, attention to time stamps, accuracy of transducers, and so on) can make a big difference to the value of the data. Conditions are often made on the use of data by ethics committees who may place great emphasis on the rights of the patients who consented explicitly for a particular research project and not other projects. This is an important subject, but can be dealt with through the process of approval and consent. From a technical perspective, the format in which data are collected may impede easy reuse and dissemination. Proprietary formats or software may require licenses and specialist knowledge. Particular datasets may also be difficult to locate. It is certainly more problematic to store accessible raw data than to make the scholarly article arising from those data available for open access.[15] Rice has outlined a variety of levels for sharing data, ranging from the "Holy-Grail of the data grid" to the "typical status quo" of the personal or networked hard drive.[15]

There does seem to be a large demand for open access data, as shown by the number of downloads from existing sets. For example, The Biomedical Signal Processing Laboratory at Portland State University, USA (http://bsp.pdx.edu/), has a signal repository for research and their website receives around 300 hits a month (http://tinyurl.com/

psbspstats). The General Practice Research Database (http://www.gprd.com) is the world's largest database of longitudinal medical records from primary care. It is made up of observational data from clinical practice. The General Practice Research Database website lists over 870 research papers published using their data. Repositories of data have been made available from intensive care[16–23] and have provided opportunities for researchers to test and develop new tools for clinical support.[24–26] However, we have been able to find few equivalent datasets of physiologic measurements made during anesthesia. One recently reported database that we did identify[27] does not include information on interventions—a critical component for understanding any physiologic changes.

In this paper we describe the establishment of two datasets containing records collected for the purposes of validating anesthetic simulators and other records collected for the purposes of validating a new diagnostic alarm. The datasets include time series of physiologic data and time-stamped information on relevant clinical interventions (such as the administration of medications). It is formatted in such a way as to facilitate extensibility so that a large-scale, searchable, open-access resource can be developed.[2 28]

METHODS

Ethics committee approval was obtained to collect anonymous data from anesthetized patients and to make the data freely available (NTX/07/16/EXP). We selected patients with no comorbidities undergoing shoulder arthroscopy. We used the SAFERSleep system (Safer Sleep LLC, Tennessee, Nashville, USA)[29] set to the highest data sampling rate (0.2 Hz; the default rate is 0.03 Hz) to collect physiologic data and information on events such as drugs administered, the surgical incision, and the insertion of an endotracheal tube: typically this information is captured manually using barcodes. Extraneous data were removed to simplify the database structure and make the data more readable and usable. Identifying data were removed to maintain confidentiality. The SAFERSleep system stores data in a proprietary XML format. We translated these raw data into a simplified XML structure (figure 1) with the support of the manufacturer and via custom R (v2.14.2 for Windows) scripts. The simplifications were designed to remove identifying data for ethical reasons, remove SAFERSleep system data (as described above), and to enhance the records with reference to standardized terms from SNOMED-CT in metadata.[30]

The same XML structure was used to store data collected from a second series of patients undergoing surgery under anesthesia in whom significant blood loss was expected (because of the nature of their surgery—most were radical prostatectomies), as part of a study into the enhancement of anesthesia alarm systems and the development of an expert diagnostic system (EDS). Again, ethics committee approval was given for this study and for publication of the data (NTX/06/08/094). These data were collected at 0.1 Hz from an AS5 anesthesia monitor (GE-Datex-Ohmeda, Helsinki, Finland). Information collected included free-form comments from clinicians about their assessments of blood volume status and other clinical indicators of blood loss, such as fluid administration and use of vasoconstrictors. Data were converted to the XML schema via the use of custom Matlab (v12) scripts.

RESULTS

The Anaesthetic Shoulder Arthroscopy Cases Dataset (http://hdl.handle.net/2292/5378) was established, and presently contains information from 13 male and 7 female patients, 21–70 years of age and weighing 57–110 kg, undergoing shoulder arthroscopy operations under general anesthesia. Each patient had 17–26 variables, 5–24 events, and 18–58 drug administrations measured during their procedure. In total, 400 physiologic data signals were collected over 20 h with 274 events and 597 drug administrations. The data collected for the EDS development (available at http://hdl.handle.net/2292/10357) contains 14 sets of physiologic data and comments. There are a total of 186 variables measured and 286 comments collected over 69 h in this dataset.

Both sets of data have been licensed with an Open Database License (http://www.opendatacommons.org/licenses/odbl), linked from the metadata accompanying each dataset. The use of the Open Database License makes the conditions of reuse clear, obviating the need for researchers to contact the creators for permission.

DISCUSSION

We have established two open access datasets of physiologic and interventional data from 34 patients undergoing surgery under anesthesia in defined conditions. The XML schema (figure 1) is sufficiently flexible to allow data to be stored from different sources, as can be seen from the two sets of data collected here. The data format is readable by humans or machines, free tools are available to manipulate the data, and it is simple to extend the dataset. XML was the format preferred to facilitate data interchange between the participants in the USA drug development process[31] and is the basis for the HL7 standard.[32] A drawback to this form of data storage is a paucity of standards for specialty-based ontologies and schemas.[4] Presently, accompanying documentation is required for interpreting the data. Other data formats were considered, including HL7 and CDA but these were considered too complicated for many possible users. This is also, presumably, the reason why similar data have previously been made available as images in addition to simple CSV files.[27]

The datasets are stored in ResearchSpace@Auckland (http://researchspace.auckland.ac.nz), the University of Auckland's institutional repository, along with a brief abstract describing the data. The repository was developed in 2006 by the University of Auckland library; the initial focus was on developing the collection of PhD theses and technical reports published by departments at the University of Auckland. A small number of theses have accompanying datasets, and this has been the catalyst for the development of a small dataset archive within the repository. The physiologic data stored in the two archives described in this paper are the only ones of their type in the ResearchSpace@Auckland repository. The present service for data is at the "zip and ship" level on the DISC-UK DataShare continuum.[15] Datasets in ResearchSpace@Auckland are discoverable as descriptive metadata harvested by Google and other search engines in a controlled manner. The use of permanent URLs for each dataset and a robust disaster recovery system reduce the risk of broken links and lost data. There are limitations to this solution; each dataset is stored individually within the wider dataset archive, but it does allow easy dissemination and reuse of data.

A limitation with the approach taken in this work is the lack of a robust quality assurance system. Potential users of the data can only rely on the authors' description of the data collection methods to make a judgment on the quality of the data. Ideally, there would be an independent review of the data before, during,

Figure 1 Structure of the XML files containing the data from recorded anesthetic cases. Bold words indicate the type of data contained between the tags.

```
<anaesthetic>
    <case>Integer</case>
    <creationtime>TimeStamp</creationtime>
    <operation>
          <opdescription SNOMED="Int">String</opdescription>
          <opdate>TimeStamp</opdate>
    </operation>
    <patient>
          <age SNOMED="105727008">Int</age>
          <weight SNOMED="27113001">Double</weight>
          <height SNOMED="50373000">Double</height>
          <sex SNOMED="263495000">String</sex>
          <dob SNOMED="263495000">DateStamp</dob>
          <asa SNOMED="Int">Int</asa>
          <comorbidities>
                <description SNOMED="Int"
                             ICD10="String">String</description>
                …
          </comorbidities>
    </patient>
    <events>
          <event>
                <evtime>TimeStamp</evtime>
                <evdescription SNOMED="Int">String</evdescription>
          </event>
          …
    </events>
    <drugs>
          <drug>
                <drname SNOMED="Int">String</drname>
                <drtime>TimeStamp</drtime>
                <drdose>
                      <drvalue>Double</drvalue>
                      <drunit>String</drunit>
                </drdose>
                <drroute>String</drroute>
          </drug>
          …
    </drugs>
    <data>
          <var>
                <vaname SNOMED="Int">String</vaname>
                <vatimes>CSV</vatimes>
                <vavalues>CSV</vavalues>
          </var>
          …
    </data>
</anaesthetic>
```

and after its collection. This would require substantial resource and limit the available data. A simpler approach would be to enable a rating system for data such that the community of users can make comments and score the dataset. Such a "reputation system" provides some level of confidence for online trading, for example,[33] and may work as a guide for data users.

One important area where community rating is inappropriate is ethics committee review. Both datasets in this paper received ethics committee approval for collection and public sharing. Ethics review boards should be encouraged to embrace the idea of sharing data,[34] provided that confidentiality of participants is assured,[35] to increase the potential value of data collected at considerable expense and sometimes some inconvenience and risk to patients. Open datasets such as the two described in this paper minimize or eliminate these risks and costs, and allow verification of the original study's analyses, and open the possibility of additional analyses.

Another limitation to the datasets described is that not all the cases contain the same sets of physiologic signals. All cases contain the commonly measured variables (heart rate, blood pressure, etc) but only some cases contain temperature, for example. This is evident from a cursory examination of the dataset and is because of the differences in clinical practice between cases. No attempt was made in the data collection to change clinical practice and so some cases have additional physiologic data.

The physiologic data were recorded directly from the anesthetic machine. However, event data required a researcher to be present as there are no automated methods of ensuring accurate recording of time and dose of drug administrations, for example. The presence of a researcher in data collection adds considerable time and cost to the creation of such datasets. These factors, combined with the removal of data identifying patients or clinicians and the potential utility of the data, were points

highlighted when applying to the ethics committee to make the data available.

CONCLUSION

We have demonstrated an approach to collecting and storing high-quality datasets from surgical operations under anesthesia. The Anesthetic Shoulder Arthroscopy Cases Dataset was used as comparison data for simulated physiology, and the dataset of records from patients expected to have blood loss was used to validate an EDS. It is conceivable that the data may be reused for other research purposes, such as pharmacodynamic model validation, simulation creation or validation, or work on diagnostic or predictive algorithms.

Acknowledgments AFM is a director of Safer Sleep, holds about 9% of its shares, and advises the company on the design of its products. The intellectual property of the new system is owned by Safer Sleep, although some patents are in the name of AFM (as inventor). AFM has been an author on several publications evaluating the SAFERSleep system.

Contributors DC collected the ASAC data, conceived the idea of a repository, converted the Anaesthetic Shoulder Arthroscopy Cases (ASAC) and expert diagnostic system (EDS) sets into the format presented, and drafted the manuscript. VN-W looked after the datasets in the repository and contributed to the manuscript. MJH collected the EDS data, contributed to the concept development, and edited the manuscript. AFM oversaw the ASAC data collection, contributed to concept development, and edited the manuscript.

Funding DC was funded by the Ralphe and Eve Seelye Scholarship in Anaesthesiology during this work.

Competing interests None.

Ethics approval Ethics approval was provided by New Zealand ethics committees.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. **Balust J,** Macario A. Can anesthesia information management systems improve quality in the surgical suite? *Curr Opin Anaesthesiol* 2009;**22**:215—22.
2. **Tremper KK.** Anesthesia information systems: developing the physiologic phenotype database. *Anesth Analg* 2005;**101**:620—1.
3. **Benson M,** Junger A, Fuchs C, *et al*. Use of an anesthesia information management system (AIMS) to evaluate the physiologic effects of hypnotic agents used to induce anesthesia. *J Clin Monit Comput* 2000;**16**:183—90.
4. **Kush RD,** Helton E, Rockhold FW, *et al*. Electronic health records, medical research, and the Tower of Babel. *N Engl J Med* 2008;**358**:1738—40.
5. **Weiss YG,** Cotev S, Drenger B, *et al*. Patient data management system in anaesthesia: an emerging technology. *Can J Anesth* 1995;**42**:914—21.
6. **Kheterpal S,** Tremper KK, Englesbe MJ, *et al*. Predictors of postoperative acute renal failure after noncardiac surgery in patients with previously normal renal function. *Anesthesiology* 2007;**107**:892—902.
7. **Chase CR,** Merz BA, Mazuzan JE. Computer Assisted Patient Evaluation (CAPE): a multi-purpose computer system for an anesthesia service. *Anesth Analg* 1983;**62**:198—206.
8. **Strauss PL,** Turndorf H. A computerized anesthesia database. *Anesth Analg* 1989;**68**:340—3.
9. **Bashein G,** Barna CR. A comprehensive computer system for anesthetic record retrieval. *Anesth Analg* 1985;**64**:425—31.
10. **Benson M,** Junger A, Fuchs C, *et al*. Using an anesthesia information management system to prove a deficit in voluntary reporting of adverse events in a quality assurance program. *J Clin Monit Comput* 2000;**16**:211—17.
11. **Reich DL,** Hossain S, Krol M, *et al*. Predictors of hypotension after induction of general anesthesia. *Anesth Analg* 2005;**101**:622—8.
12. **Cumin D,** Weller JM, Henderson K, *et al*. Standards for simulation in anaesthesia: creating confidence in the tools. *Br J Anaesth* 2010;**105**:45—51.
13. **Wheeler D,** Ward C, Jukes M, *et al*. *JISC FInal Report: UK Colonial Registers and Royal Navy Logbooks*. University of Sunderland, 2009.
14. **National Institutes of Health.** *Final NIH Statement on Sharing Research Data*. 2003. http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html (accessed 2 Dec 2008).
15. **Rice R.** DISC-UK DataShare. *ALISS Q* 2008;**3**:7—12.
16. **Norris PR,** Riordan WP, Dawant BM, *et al*. SIMON: a decade of physiological data research and development in trauma intensive care. *J Healthc Eng* 2010;**1**:315—35.
17. **Crichton DJ,** Mattmann CA, Hart AF, *et al*. An informatics architecture for the virtual pediatric intensive care unit. *International Symposium on Computer-Based Medical Systems; 2011*. 2011.
18. **Saeed M,** Lieu C, Raber G, *et al*. MIMIC II: a massive temporal ICU patient database to support research in intelligent patient monitoring. *Comput Cardiol* 2002;**29**:641—4.
19. **Kovács S,** McQueen DM, Peskin CS. Modelling cardiac fluid dynamics and diastolic function. *Phil Trans Roy Soc Lond* 2001;**359**:1299—314.
20. **Goldberger AL,** Amaral LA, Glass L, *et al*. Physiobank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation* 2000;**101**:215—20.
21. **Gade J,** Korhonen I, Van Gils MJ, *et al*. Technical description of the IBIS data library. *Comput Methods Programs Biomed* 2000;**63**:175—86.
22. **Korhonen I,** Ojaniemi J, Nieminen K, *et al*. *Building the IMPROVE Data Library*. IEEE Engineering in Medicine and Biology, 1997:25—32.
23. **Hart GK.** The ANZICS CORE: an evolution in registry activities for intensive care in Australia and New Zealand. *Crti Care Resusc* 2008;**10**:83—8.
24. **Rocha T,** Paredes S, de Carvalho P, *et al*. Prediction of acute hypotensive episodes by means of neural network multi-models. *Comput Biol Med* 2011;**41**:881—90.
25. **McBride J,** Sullivan A, Xia H, *et al*. Reconstruction of physiological signals using iterative retraining and accumulated averaging of neural network models. *Physiol Meas* 2011;**32**:661—75.
26. **Burykin A,** Peck T, Krejci V, *et al*. Toward optimal display of physiologic status in critical care: I. Recreating bedside displays from archived physiologic data. *J Crit Care* 2011;**26**:105.e1—9.
27. **Liu D,** Görges M, Jenkins SA. Development of an accessible repository of anesthesia patient monitoring data for research. *Anesth Analg* 2012;**114**:584—9.
28. **Bierstein K.** Anesthesia information systems. where awareness is good! *American Society of Anesthesiologists Newsletter* 2007;**71**:37—9.
29. **Merry AF,** Webster CS, Mathew DJ. A new, safety-oriented, integrated drug administration and automated anesthesia record system. *Anesth Analg* 2001;**93**:385—90.
30. **IHTSDO.** *International Health Terminiology Standards Development Organization*. http://www.ihtsdo.org/ (accessed 4 Jun 2012).
31. **Canfield K.** Improving interorganizational data interchange for drug development. *Comput Biol Med* 1999;**29**:89—99.
32. **Orgun B,** Vu J. HL7 ontology and mobile agents for interoperability in heterogeneous medical information systems. *Comput Biol Med* 2006;**36**:817—36.
33. **Carbral L,** Hortaçsu A. The dynamics of seller reputation: evidence from ebay. *J Ind Econ* 2010;**58**:54—78.
34. **Sieber JE.** Sharing scientific data I: new problems for IRBs. *Ethics Hum Res* 1989;**11**:4—7.
35. **Anderson BJ,** Merry AF. Data sharing for pharmacokinetic studies. *Pediatr Anesth* 2009;**19**:1005—10.