BT-Nurse: Computer Generation of Natural Language Shift Summaries from Complex Heterogeneous Medical Data

1. James Hunter, Department of Computing Science, University of Aberdeen, DPhil

2. Yvonne Freer, Simpson Centre for Reproductive Health, Royal Infirmary of Edinburgh, PhD

3. Albert Gatt, Institute of Linguistics, Centre for Communication Technology, University of Malta, PhD

4. Ehud Reiter, Department of Computing Science, University of Aberdeen, PhD

5. Somayajulu Sripada, Department of Computing Science, University of Aberdeen, PhD

6. Cindy Sykes, Simpson Centre for Reproductive Health, Royal Infirmary of Edinburgh

7. Dave Westwater, Data2Text, Aberdeen, MSc

Corresponding author:
Prof J Hunter
Mailing address: Department of Computing Science, University of Aberdeen, King's College, Aberdeen, AB24 3UE, UK
Tel: +44 1224 272287
FAX: +44 1224 273422
email: j.hunter@abdn.ac.uk

I, as corresponding author, promise that I and all persons listed as coauthors on this submitted work have read and understand the "Originality of Manuscripts statement regarding submissions to JAMIA (available at http://jamia.bmj.com/site/about/originalityofmanuscripts.xhtml and confirm that this submission is a new, original work that has not been previously submitted, published in whole or in part, or simultaneously submitted for publication in another journal. Also, in accordance with the aforementioned policy, we have included as part of the submission any previously published materials that overlap in content with this new original manuscript.

The Corresponding Author has the right to grant on behalf of all authors and does grant on behalf of all authors, an exclusive license (or non-exclusive for government employees) on a worldwide basis to BOTH The American Medical Informatics Association and its publisher for JAMIA, the BMJ Publishing Group Ltd and its Licensees to permit this article (if accepted) to be published in Journal of the American Medical Informatics Association and any other BMJPGL products to exploit all subsidiary rights, as set out in our license (http://group.bmj.com/products/journals/instructions-for-authors/licence-forms).

Abstract

Objective: To determine if a computer system can automatically generate a useful natural language nursing shift summary solely from an electronic patient record system, in a neonatal intensive care unit (NICU).
Design: A system was built which automatically generates NICU shift summaries, using data-to-text technology. The system was tested for two months in the Royal Infirmary of Edinburgh NICU.
Measurements: Nurses were asked to rate the understandability, accuracy, and helpfulness of the computer-generated summaries; they were also asked for free-text comments about the summaries.
Results: The nurses found the majority of the summaries to be understandable, accurate, and helpful ($p < .001$ for all measures). However, nurses also pointed out many deficiencies, especially with regard to extra content they wanted to see in the computer-generated summaries.
Conclusions: Natural language NICU shift summaries can be automatically generated from an electronic patient record. However our proof-of-concept software needs considerable additional development work.

I. Introduction

Medical professionals have access to increasing volumes of information about patients. This is particularly the case in the ICU, where continuously monitored physiological data (e.g. heart rate, oxygen saturation) and detailed records of observations and interventions are available.

Effective presentation and understanding of these data is important in real-time decision making, but is also very relevant during patient handover between clinicians where, despite an oral or written handover, the outgoing clinician may forget to mention important information, or the incoming clinician may not assimilate all the information presented during a brief exchange. In such cases the incoming clinician relies on the available data in order to fill in any gaps.

While large data sets are usually presented as graphs or tables, some studies have found that high-quality textual summaries can be more effective for decision-support in some circumstances [1,2]. The summaries in these studies were carefully written by expert clinicians; this is only practical within a research context. However, as part of a larger project, BabyTalk [3,4], we have developed a Natural Language Generation system, BT-Nurse, which *automatically* generates English summaries of the electronically recorded patient data over a twelve hour nursing shift, for a baby in a neonatal ICU (NICU).

[Figure 1 about here.]

II. Case Description

The NICU at the Edinburgh Royal Infirmary uses Clevermed®'s *Badger* computer system to manage and display patient data. This system acquires and records several channels of continuous physiological data sampled once per second. A display is located beside each cot, and clinical staff routinely use this to enter additional information, including hourly physiological measurements, drugs and fluids administered, equipment settings, care and treatment actions taken, etc. Most of the data collected is pre-formatted but free-text entry is also available.

BT-Nurse analyses the patient data, decides which information is most important and presents it as an English text; an extract is shown in Figure 1a. Figure 1b shows a corresponding extract from a summary of the same shift data written by a research nurse. The summaries are structured according to physiological system (e.g. *respiratory*); BT-Nurse only summarises two of the ten physiological systems, viz. *respiratory* and *cardiovascular*; it also reports on the patient's current status and problems.

III. Methods of Implementation

BT-Nurse is constructed around a standard data-to-text 'pipeline' architecture (Figure 2) [5,6] where information is processed sequentially by modules which communicate via a domain ontology which includes mechanisms for modeling uncertainty and incomplete knowledge.

[Figure 2 about here.]

1. *Data Translation* transforms data from the format stored in the Badger system to that required by the ontology. A limited amount of information is extracted from free-text, using simple parsing and keyword extraction techniques. For example, the text *bolus of sodium chloride infusing* would be mapped to an instance of a *drug administration* event in the ontology, with properties indicating the drug (sodium chloride) and the method (infusion).
2. *Data Pre-Processing* tries to fill in some of the omissions and gaps which are inevitably present in real world patient data. For example, if the time of an intubation has not been entered, BT-Nurse attempts to infer approximately when it happened by examining the hourly ventilation observations.
3. *Signal Analysis* detects and removes artifacts from the physiological data and extracts a small number of events, both short-term (e.g. bradycardias and desaturations) and long-term (e.g. trends and abstractions such as "within normal range"). Figure 3 presents the data corresponding to the summaries in Figure1.

[Figure 3 about here.]

4. *Data Interpretation* uses medical knowledge to enhance the information recorded in the ontology (i) by estimating the medical significance of events, (ii) by deriving higher-level *abstractions* (e.g. the *state* of having respiratory acidosis), and (iii) by inferring causal and other relationships between events. Medical knowledge is expressed using forward chaining rules derived during extensive knowledge acquisition exercises with domain experts (a consultant neonatologist and senior neonatal nurse).
5. *Document Planning* decides on the content and structure of the generated text. Some sections of the text, such as *Current Status* (see Figure 1a), essentially have fixed structures which are populated by relevant events. For other sections, such as *Events During the Shift*, the document planner identifies a small number of key events during the shift (based on medical significance) and generates a paragraph around each of these.
6. *Microplanning and Realisation* maps the ontology instances selected by the Document Planner to English text by (i) mapping each ontology instance to a semantic structure using rules which select linguistic predicates; (ii) aggregating the resulting representations into higher-order structures; (iii) realising the structures as English text using the SimpleNLG realization engine [7], which performs English syntactic and morphological generation.

IV. Example and Observations

BT-Nurse was evaluated on-ward by nurses who read summaries about babies under their care. These summaries were constructed on demand towards the end of selected shifts using the live database and displayed to the nurse at the cot-side. They were generated in less than one minute, with no noticeable impact on other users of the Badger system. *No* additional data-entry was required; all information was obtained from the Badger patient record.

After reading a summary, the *outgoing* nurse in charge of that specific baby was asked to rate its *understandability, accuracy* and *helpfulness in producing her own end-of-shift summary* (were she to produce one), by indicating *agreement*, *disagreement* or *neutrality* with respect to these statements. She was also invited to enter additional comments on any aspect of the summary. The protocol for *incoming* nurses was the same, except that the final question referred to *helpfulness in care planning*. Incoming nurses had a verbal handover with the outgoing nurse and also had access to the computerized charts; hence they could judge the accuracy and helpfulness of a generated summary. We could not directly compare BT-Nurse texts against human-written texts, using measures such as recall and precision, because NICU nurses do not write detailed textual shift handover reports.

We conducted 165 trials (defined as an evaluation by one nurse of one shift summary): 73 with outgoing and 92 with incoming nurses. In 131 cases, a summary was seen by only one nurse (outgoing or incoming); in the other 17 cases the summary was seen by both nurses. A total of 148 summaries were produced for 31 individual babies. On average, each baby was seen by 2.3 nurses (maximum 6). Of a nursing staff complement of 93, 54 different nurses participated. On average each nurse saw 4.0 different babies; only 4 nurses saw more than 8 different babies.

We compared response frequencies for each of the categories in each of the three questions that nurses were asked, using a χ^2 test on response frequencies by items (i.e. over trials). Overall, there were significant differences between the number of positive, negative and neutral responses for *understandability* ($\chi^2 = 241.89$; $p < .001$), *accuracy* ($\chi^2 = 110.22$; $p < .001$), and *helpfulness* ($\chi^2 = 64.15$; $p < .001$). As Table 1 shows, the majority response was positive in all three cases. A multinomial logistic regression showed no no significant differences between incoming and outgoing nurses for any of these questions (*understandability:* model $\chi^2 = 15.26$, $p = .08$; *accuracy:* $\chi^2 = 19.99$; $p = .1$; *helpfulness* $\chi^2 = 17.29$; $p > .7$).

	Understandability			**Accuracy**			**Helpfulness**		
	Agree	**Neutral**	**Disagree**	**Agree**	**Neutral**	**Disagree**	**Agree**	**Neutral**	**Disagree**
Incoming	92.4	7.6	0	73.9	23.9	2.2	56.5	35.9	7.6
Outgoing	87.7	8.2	4.1	65.8	24.7	9.6	61.6	30.1	8.2
Overall	90.3	7.9	1.8	70.3	24.2	5.5	58.8	33.3	7.9

Table 1: Nurses' views of the BT-Nurse summaries (% of trials)

The comments were manually segmented, so that each segment addressed one specific aspect of the summary. This yielded 237 segments (125 for outgoing nurses and 112 for incoming). The segments were annotated independently by three of the authors to indicate which aspect of a summary each was concerned with (*content*, *language* and *overall*) and which of a predefined set of labels for each category applied. We used Cohen's κ statistic to calculate pairwise agreement between annotators on each dimension; using standard thresholds [8], we found *tentative* agreement in the *content* ($\kappa = 0.73$) and *language* ($\kappa = 0.66$) dimensions and *good* agreement in the *overall* ($\kappa = 0.83$) dimension.

Most segments concerned the *content* of the summary (185), most of these (109) noting *missing content*. Many of these referred to information which was not intended to be included (e.g. content about nutrition, which BT-Nurse did not handle, as noted in Section II). However, even disregarding such segments, requests for more content were much more common than requests for less, suggesting that BT-Nurse was under-reporting. As an example, one nurse wrote that the *baby […] is VERY small, and […] it should be pointed out that the ETT is size 2.0.* BT-Nurse never reports ETT size as usually this is not very important; however in some cases it is important and should be reported.

There were 46 segments concerning *incorrect content*; some were due to errors in the patient record data, but most were due to bugs in the software. For example, BT-Nurse sometimes listed *current problems* which in fact had been rectified; this was because of an error in reading the relevant database table.

There were only 11 segments about language, all of which were negative. Most of these criticisms reflected individual preferences. The small number of such segments raises the possibility that nurses only commented about language when they were unhappy with it; if this is true, then perhaps in most cases BT-Nurse's language, which is stylistically quite different from the somewhat telegraphic style that typifies free-text comments in normal shift summaries, was "good enough".

Among the 35 *overall* segments 8 concerned deficiencies from a high-level "narrative" perspective - for example, not describing causal links between observations and interventions, and not adequately describing the overall "big picture."

There were also some very encouraging comments about how BT-Nurse summaries were helpful, such as *BT picked up the change in HR trend that I had not noticed.*

V. Discussion

Given that we did not have time to develop a system which generated complete shift summaries, we think it is very encouraging that 58% of the nurses regarded BT-Nurse texts as helpful, and 90% regarded them as understandable.

Most of the criticisms do not concern the underlying technology and could be addressed by expanding BT-Nurse so that it generates complete summaries (including systems that were omitted, such as nutrition), and by doing more on-ward debugging. In terms of technology, the biggest challenges are dealing with incomplete input data and generating good narrative texts. Data entered manually will always have omissions and mistakes and dealing with these robustly is a major challenge for any medical data-to-text system. Generating good narratives which include causal links and make the big picture clear is also a key data-to-text challenge; indeed, one could argue that such narrative aspects are perhaps the primary benefits of textual summaries over tabular/visual presentations. BT-Nurse's medical knowledge base also needs to be expanded.

Data-to-text technology is very new, and systems have been developed in many areas, including weather forecasts [9,10,11,12,13], financial and statistical information [14,15,16], and engineering [17]. Recent attempts to apply data-to-text in medical contexts [18,19] have used input which is simple compared with BT-Nurse; these are akin to early automated interpretation and report generation systems for personality assessment based on questionnaire responses [20]. We are not aware of any previous medical system which is as ambitious as BT-Nurse in the amount and diversity of data summarized.

BT-Nurse has shown that it is possible to use data-to-text technology to generate useful and helpful summaries of nursing shifts from a complex, state-of-the-art patient information system holding a large amount of heterogeneous data. Of course, BT-Nurse is a proof of concept, and would require considerable engineering effort before it could be realistically deployed, or indeed evaluated in a clinical trial which measured patient outcome instead of nurse's perceptions. In particular, report accuracy needs to be higher, and comparable to the overall accuracy of the information in the patient record system. However, our evaluation, which involved a deployment of the system within its target environment and running on live, previously unseen data, shows that data-to-text systems can generate shift summaries from clinical data extracted from an electronic patient record.

Acknowledgements

We are grateful to the UK Engineering and Physical Sciences Research Council (EPSRC) for funding the BabyTalk project with grants to the University of Aberdeen (EP/D049520/1) and the University of Edinburgh (EP/D05057X/1).

References

1. Law AS, Freer Y, Hunter JRW, Logie RH, McIntosh N, Quinn J. A comparison of graphical and textual presentations of time series data to support medical decision making in the neonatal intensive care unit. Journal of Clinical Monitoring and Computing. 2005; 19: p. 183-194.
2. van der Meulen M, Logie RH, Freer Y, Sykes C, McIntosh N, Hunter JRW. When a graph is poorer than 100 words: A comparison of computerised Natural Language Generation, human generated descriptions and graphical displays in neonatal intensive care. Applied Cognitive Psychology. 2008; 24: p. 77-89.
3. Gatt A, Portet F, Reiter E, Hunter JRW, Mahamood S, Moncur W, et al. From data to text in the neonatal intensive care unit: Using NLG technology for decision support and information management. AI Communications. 2009; 22: p. 153-186.
4. Portet F, Reiter E, Gatt A, Hunter JRW, Sripada S, Freer Y, et al. Automatic generation of textual summaries from neonatal intensive care data. Artificial Intelligence. 2009; 173: p. 789-816.
5. Reiter E, Dale R. Building Natural Language Generation Systems Cambridge: Cambridge University Press; 2000.
6. Reiter E. An architecture for data-to-text systems. In Buseman S, editor. Proceedings of the 11th European Workshop on Natural Language Generation; 2007; Schloss Dagstuhl, Germany: ACL. p. 97-104.
7. Gatt A, Reiter E. SimpleNLG: A realisation engine for practical applications. In Krahmer E, Theune M, editors. Twelfth European Workshop on Natural Language Generation: Proceedings of the Workshop; 2009; Athens, Greece: ACL. p. 90-94.
8. Artstein R, Poesio M. Inter-coder agreement for computational linguistics. Computational Linguistics. 2008; 34(4): p. 555-596.
9. Goldberg E, Driedger N, Kittredge RI. Using Natural Language Processing to produce weather forecasts. IEEE Expert. 1994; 9: p. 45-53.
10. Coch J. MULTIMETEO: Multilingual Generation of Weather Forecasts. ELRA Newsletter. 1998; 3(2).
11. Reiter E, Sripada S, Hunter JRW, Yu J, Davy I. Choosing words in computer-generated weather forecasts. Artificial Intelligence. 2005; 167: p. 137-169.
12. Turner R, Sripada S, Reiter E, Davy I. Selecting the content of textual descriptions of geographically located events in spatio-temporal weather data. In Ellis R, Allen T, Petridis M, editors. Applications and Innovations in Intelligent Systems XV. Proceedings of the 27th SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence.; 2007; Berlin: Springer. p. 75-88.
13. Belz A. Automatic generation of weather forecast texts using comprehensive probabilistic generation space models. Natural Language Engineering. 2007; 14: p. 431-455.

14. Kukich K. Design of a Knowledge-Based Report Generator. In Marcus M, editor. Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics; 1983; Cambridge, MA: ACL. p. 145-150.
15. Iordanskaja L, Kim M, Kittredge R, Lavoie B, Polguere A. Generation of extended bilingual statistical reports. In Proceedings of the 15th International Conference on Computational Linguistics; 1992; Nantes: ICCL. p. 1019-1023.
16. Ferres L, Parush A, Roberts S, Lindgaard G. Helping people with visual impairments gain access to graphical information through natural language: The iGraph system. In Meisenberger K, Klaus J, Zagler W, Karshler A, editors. Computers Helping People with Special Needs: 10th International Conference ICCHP 2006; 2006; Berlin: Springer. p. 1122-1130.
17. Yu J, Reiter E, Hunter JRW, Mellish C. Choosing the content of textual summaries of large time-series data sets. Natural Language Engineering. 2007; 13: p. 25-49.
18. Dalal M, Feiner S, McKeown K, Jordan D, Allen B, al Safadi Y. MAGIC: An experimental system for generating multimedia briefings about post-bypass patient status. In Cimino JJ, editor. Proceedings of the AMIA Annual Fall Symposium; 1996; Washington, DC: Henley and Belfus Inc. p. 684-688.
19. Harris M. Building a large-scale commercial NLG system for an EMR. In White M, Nakatsu C, McDonald D, editors. INLG 2008: Fifth International Natural Language Generation Conference; 2008; Salt Fork, Ohio: ACL. p. 157-160.
20. Moreland KL. Validation of computer-based test interpretations: Problems and prospects. Journal of Consulting and Clinical Psychology. 1985; 53(6): p. 816-825.

Figures

Figure 1: Extracts from nursing shift summaries: (a) Generated by BT-Nurse; (b) Written by a research nurse.

Figure 2: BT-Nurse architecture

Figure 3: Example physiological data sampled once per second

(a)

Respiratory Support

Current Status

Currently, the baby is on CMV in 27 % O2. Vent RR is 55 breaths per minute. Pressures are 20/4 cms H2O. Tidal volume is 1.5.

SaO2 is variable within the acceptable range and there have been some desaturations.

The most recent blood gas was taken at around 07:45. Parameters are acceptable. pH is 7.3. CO2 is 5.72 kPa. BE is -4.6 mmol/L. The last ET suction was done at about 05:15.

Events During the Shift

A blood gas was taken at around 19:45. Parameters were acceptable. pH was 7.18. CO2 was 7.71 kPa. BE was -4.8 mmol/L.

Another ABG was taken at around 23:00. Blood gas parameters had deteriorated to respiratory acidosis by around 23:00. pH was 7.18. CO2 had risen to 9.27 kPa by around 23:00. BE was -4.8 mmol/L.

The baby was intubated at 00:15 and was on CMV. Vent RR was 50 breaths per minute. Pressures were 20/4 cms H2O. FiO2 was 29 %. Tidal volume was 1.5. He was given morphine and suxamethonium. MAP was raised from 6 cms H2O to 8 cms H2O.

Between 00:30 and 03:15, SaO2 increased from 88 % to 97 %.

Another ABG was taken at around 00:45. pH was 7.18. CO2 dropped to 7.95 kPa. BE was -4.8 mmol/L.

...

(b)

Respiratory support needed due to immaturity

CURRENT MANAGEMENT / ASSESSMENT:

CMV rate 55, pressures 20/4, in 27% oxygen, giving tidal volumes of 1.5 ml (3.3 ml/kg). Very recent ABG good: pH 7.31, CO_2 5.72. He received morphine prior to intubation at 00:30; no spontaneous respiratory effort noted since being re-ventilated. Desaturates during cares and suction but recovers afterwards; otherwise SpO_2 has been fairly stable. Large ETT secretions, mucopurulent and blood stained.

EVENTS DURING THIS SHIFT:

While on BiPAP, oxygen requirement increased to 50% by 23:00

ABG at 23:10 showed CO_2 increased from 7.7 to 9.27 in three hours

Electively re-intubated at 00:30 to CMV rate 50, pressures 18/4 in 30% oxygen

Difficult intubation; size 2 ETT

Was given morphine and suxamethonium

On ventilation, oxygen requirement reduced to 30% and ABG initially improved

...

Figure 1: Extracts from nursing shift summaries:
(a) Generated by BT-Nurse; (b) Written by a research nurse.

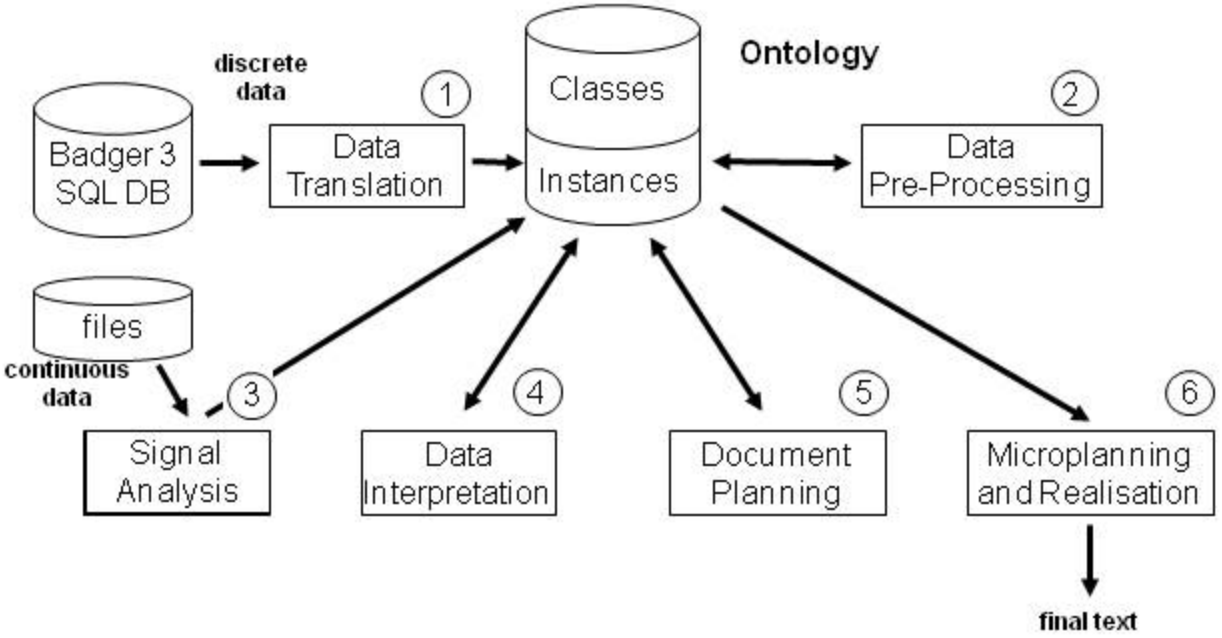

Figure 2: BT-Nurse architecture

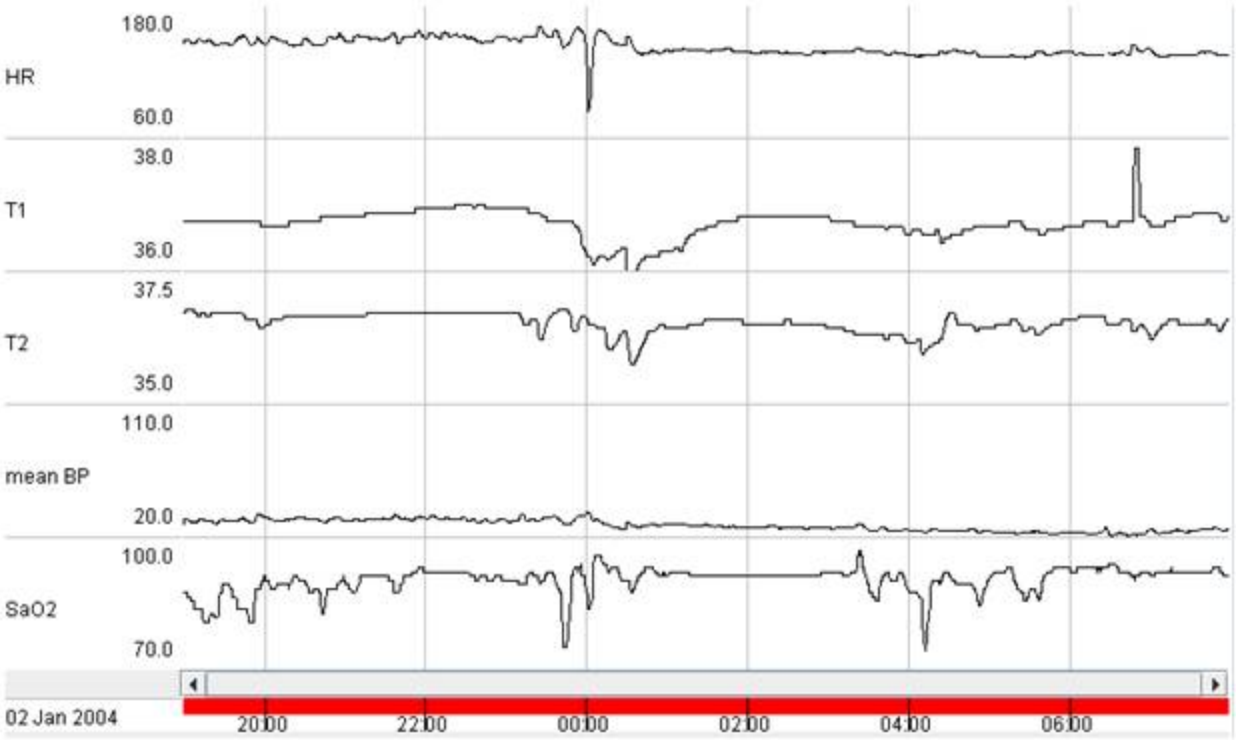

Figure 3: Example physiological data sampled once per second