

DATASHEET:

African Surgical Sheet Dataset

Ryan D. Folks
University of Virginia Department of
Anesthesiology
vcz2aj@uvahealth.org

Bhiken I. Naik
University of Virginia Department of
Anesthesiology
bin4n@uvahealth.org

This document is based on *Datasheets for Datasets* by Gebru *et al.* [1]. Please see the most updated version [here](#).

MOTIVATION

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

This dataset was created to capture a pilot sample of surgical data from low and middle income countries in sub-Saharan Africa, a place where digitized surgical data is not widely available yet.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

This dataset was created by the University of Virginia and Safe Surgery South Africa on behalf of the Lacuna Fund.

What support was needed to make this dataset? (e.g. who funded the creation of the dataset? If there is an associated grant, provide the name of the grantor and the grant name and number, or if it was supported by a company or government agency, give those details.)

The Lacuna Fund provided the funding for the creation of this dataset.

Any other comments?

A secondary motivation for the creation of this dataset was to collect image data (not available as a part of this dataset) to serve as training data for computer vision software that can digitize surgical sheets automatically as a way of bridging the gap between having no electronic medical records, and having the full infrastructure for totally digital medical records.

COMPOSITION

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes

and edges)? Please provide a description.

Instances in our dataset represent all of the data recorded for a single surgery available on a paper surgical record. This includes demographics, vitals, lab values, checklists, hemodynamics, physiological indicators, and surgical metadata. This data is indexed with a single key across multiple tables.

How many instances are there in total (of each type, if appropriate)?

There are 2,664 surgeries in total.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

Our dataset is a convenience sample of surgeries across a one year time span. The surgeries in this dataset do not necessarily comprise a representative sample of the kinds of surgeries performed at each site, nor do they represent the surgical population in general.

What data does each instance consist of? “Raw” data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each surgery (instance) is comprised of the patient’s demographic data (age, gender, weight (kg), height (cm), and bmi), preoperative vitals (systolic blood pressure, diastolic blood pressure, heart rate, and oxygen saturation), asa status, whether or not the surgery is emergent or elective, laboratory values (varies between sites but may include Hb, MCV, MCH, Platelets, WCC, Na, K, Urea, Creatinine, Cl, Ca, Mg, PO4, ALT, AST, ALP, Albumin, aPTT, INR, and Glycemia), checkboxes encoding binary information concerning the surgery (anesthesia type, airway management, monitoring details, postoperative pain management, and postoperative disposition location), intraoperative hemodynamic data (systolic and diastolic blood pressure as well as heart rate recorded every

5 minutes), intraoperative physiological data (oxygen saturation, fraction of inspired oxygen, end tidal carbon dioxide, tidal volume and respiratory rate if a ventilator is being used), and which drugs were given during the procedure. Additional data is available on a site by site basis.

Is there a label or target associated with each instance? If so, please provide a description.

The number of days spent in hospital post surgery as well as in-hospital death are recorded for each patient and are the major postoperative outcomes recorded for this dataset.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

Data for each patient is randomly missing due to the data itself not being recorded on the physical chart, or in some rare instances being illegible.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

There are no recorded relationships between instances.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

There are no recommended data splits.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

Data was transcribed by hand, so errors in transcription will occur infrequently.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The data is self-contained.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

This dataset includes patient data from surgeries. Site 1 contains data on what surgery is being performed. Because of this, all observations have been totally de-identified, and include no dating or time of day information.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

No.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Yes, this data set directly relates to people.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

Our data set records both gender and age. Age is binned into 10 year bins (Eg: 0-10, 50-60)

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

Every surgery has its date, mrn, and time of day removed. In addition, sites are given identifiers (SITE1, SITE2), which further obfuscate the specifics of the surgery, so it is highly unlikely that a patient could be identified based on this information.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

This data set is made up of health data, which makes it sensitive data.

Any other comments?

No

COLLECTION

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

This dataset was directly acquired by the human transcription

of photographs of paper surgical records.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created. Finally, list when the dataset was first published.

The data was collected over the course of one year in the early 2020s.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

The paper charts were filled out by anesthesiologists at each site, then photographed by clinical research coordinators on site. The photographs were securely uploaded and transcribed by the University of Virginia.

What was the resource cost of collecting the data? (e.g. what were the required computational resources, and the associated financial costs, and energy consumption - estimate the carbon footprint. See Strubell *et al.*[2] for approaches in this area.)

This dataset cost 500 USD for a smartphone and 20,000 USD for data collectors.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

The charts collected were a convenience sample of the surgeries performed at each site. The research coordinators were limited by how many patients they could follow up on during a given time frame in order to collect post-operative outcomes data.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

The clinical research coordinators were paid through grant money from the Lacuna Fund to collect outcomes data and photograph paper surgical charts. Data entry professionals at the University of Virginia are on payroll of the Department of Anesthesiology, and were thereby compensated.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

IRB approval was obtained from all sites prior to data collection.

Does the dataset relate to people? If not, you may skip the remainder of the questions in this section.

This data set directly relates to people and the surgeries they underwent.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

The data was obtained directly.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

Yes, consent was obtained from each individual.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

Yes, consent was obtained from each individual.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate)

Yes, consenting individuals can revoke their consent at any time.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

No.

Any other comments?

None

PREPROCESSING / CLEANING / LABELING

Was any preprocessing/cleaning/labeling of the data done(e.g.,discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

No preprocessing was performed. The data is presented in its final form as a digitized version of the paper intraoperative chart.

Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the “raw” data.

The "raw" data, in this case the photographs of the paper

charts, was saved, but is not available for distribution, but is available for cross-reference by those allowed to view the raw data in the event a probable mistake is caught.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

This software is not publically available.

Any other comments?

None.

USES

Has the dataset been used for any tasks already? If so, please provide a description.

No.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

Yes, in the future.

What (other) tasks could the dataset be used for?

Intraoperative anesthesia practice patterns, timeseries analysis, and postoperative outcomes research.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

The dataset was transcribed by non-medical professionals, so mistakes in transcription may result in values that are not physiologically plausible.

Are there tasks for which the dataset should not be used? If so, please provide a description.

No.

Any other comments?

None

DISTRIBUTION

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

Yes, it is available to any institution or individual contingent upon a data use agreement.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

The dataset will be available through the iThriv portal upon approval of a data use agreement.

When will the dataset be distributed?

No later than January 2025.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions. There will be a data use agreement associated with the use of this dataset outlining the proper use and distribution of it.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No.

Any other comments?

None.

MAINTENANCE

Who is supporting/hosting/maintaining the dataset?

The University of Virginia department of Anesthesiology in conjunction with iThriv will host and maintain this dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

The owners of this dataset can be contacted via email.

Is there an erratum? If so, please provide a link or other access point.

No.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

Not all charts photographed have been transcribed and will therefore not be immediately available upon the release of this dataset. The dataset will eventually be updated with all the chart data transcribed.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

No, the data will be available for use in perpetuity.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

No, only the newest version will be available at any point in time.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

Not as yet.

Any other comments?

None

REFERENCES

- [1] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumeé III, and Kate Crawford. Datasheets for Datasets. *arXiv:1803.09010 [cs]*, January 2020.
- [2] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and Policy Considerations for Deep Learning in NLP. *arXiv:1906.02243 [cs]*, June 2019.