Big data methods for CT doses analysis and monitoring

Pierre-Luc Asselin[1], Yannick Lemaréchal[2], Gabriel Couture[2], Samuel Ouellet[1], Jonathan Boivin[3], and Philippe Després[1,2,3]

[1]Département de physique, de génie physique et d'optique, Université Laval, Québec, Québec, Canada.
[2]Centre de recherche sur le cancer, Université Laval, Québec, Québec, Canada.
[3]Service de physique médicale et de radioprotection, CHU de Québec–Université Laval, Québec, Québec, Canada.

Abstract
Computed Tomography is the principal contributor to ionizing radiation exposure of medical origin for the Canadian population. The Canadian Computed Tomography Survey by Health Canada provides guidance in the safe use of radiation-emitting devices, notably by proposing Diagnostic Reference Levels (DRLs) for typical CT exams (e.g. head, chest, abdomen). Dose associated with typical studies can be compared to DRLs, but this remains a periodic one-off procedure that, although part of a quality assurance program, might fail to capture the reality of clinical activities. There is a need to monitor more closely and continuously the use of ionizing radiation in CT, as increasingly demanded by regulatory authorities. This project aims to implement big data methods to provide CT dose visualization and analytic tools to managers and technical personnel in a large institution. For instance, dashboards were deployed to provide an overview of radiation usage in CT, leading to opportunities to study clinical practice trends and facilitating the identification of outliers to improve heath care. The technology used was found to be well adapted to the analysis of massive datasets without interfering with clinical activities.

1 Introduction

Computed Tomography is currently the principal contributor to ionizing radiation exposure of medical origin for the Canadian population [1], a situation that typically prevails in developed and developing countries. CT examination requests are also on the rise, with an increase of 30% from 2010-2020 and an additional expected increase of 18% within the next twenty years [2]. The Canadian Computed Tomography Survey by Health Canada provides guidance in the safe use of radiation-emitting devices, notably by proposing Diagnostic Reference Levels (DRLs) for typical CT exams (e.g. head, chest, abdomen). Further dosimetric recommendations are also given by the Safety Code 35 to promote better radiation exposition management within health institutions [3]. Proper quality assurance programs respect such guidelines, but might still fail to capture the reality of clinical activities.

In 2016, authorities in Quebec (Canada) surveyed current dosimetric practices within the province and demanded both to better conform to Health Canada guidelines, and to monitor more closely the use of ionizing radiation in CT. Currently, no clear guidelines or best practices are proposed to health institutions in Québec for radiation exposure monitoring. Therefore, this project aims at developing a dashboard to provide visualization and analytic tools to managers and technical personnel to monitor radiation usage in CT.

2 Methods

2.1 DICOM

All developments were aligned with the Digital Imaging and Communications in Medicine (DICOM) standard, to maximize semantic and technical interoperability across workflow components. The DICOM standard is well established for the communication and the storage of medical imaging, clinical data and associated metadata. DICOM's standardized format ensures a robust and predictable structure, unlike more generic formats such as XML and JSON. DICOM is a strong foundation to build on, especially with FAIR data management principles in mind.

The DICOM standard provides more than 3600 data fields to store data and contextual information. For instance, a DICOM object could contain information such as the age of the patient, along with weight, date of birth and sex. DICOM data files can be seen as collections of standardized data fields related to a shared subject, with unique identifiers for sound data management. These files can then reference each other to better represent the complex relations between them, as expected in a medical imaging context. For example, a DICOM data file containing the actual image could reference other files pertaining to the patient or study such as segmentations, dose maps or dose reports.

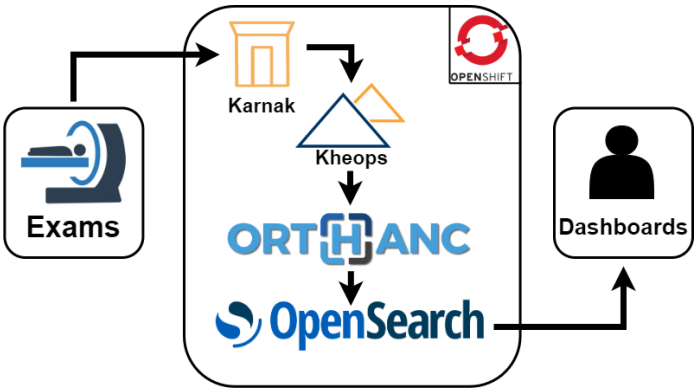

Figure 1: General flow within the PARADIM platform to expose data to OpenSearch.

2.2 Data flow

CT dose monitoring dashboards are fed by a data flow generated by routine clinical activities. Software components supporting this data flow are hosted at Université Laval within our custom-made PARADIM platform (*Plateforme d'annotation, de réutilisation et d'analyse d'images médicales* - Medical image annotation, reuse and analysis platform)[1]. PARADIM includes these components:

- Clinical data extraction and de-identification (`Karnak`)
- Data governance (auth/autz) (`Kheops`)
- Data storage (`Orthanc`)
- Data viz environment (`OpenSearch`)

`Karnak` - `Karnak` is an open-source DICOM gateway for data de-identification and DICOM attribute normalization. This software acts as a DICOM listener connected to the PACS (*Picture Archiving and Communication System*) and feeds de-identified data to Kheops through DICOMWeb. Data de-identification is supported field by field and includes: replacement by a dummy value, time-shift for dates or outright field removal. Various de-identification profiles can be managed within Karnak, including the DICOM basic privacy profile[4].
`Kheops` - `Kheops` is a DICOM-compliant image archive under MIT license offering data governance functionalities, including granular authentification/authorization (auth/autz) for data sharing for human an machine as well (through revokable API tokens). Images within `Kheops` are managed as data collections (called albums) accessible only with proper rights. Read and write permissions, as well as download and sharing rights are granted separately. A collection within Kheops was constituted for this particular project.
`Orthanc` - `Orthanc` is an open-source DICOM server used as the main DICOM database in PARADIM, with S3-compatible Ceph (RedHat) as an object storage backend. `Orthanc` provides a graphical web interface and RESTful API to request DICOM data. Moreover, `Orthanc` includes a Python (or lua) code interpreter to handle/transform DICOM data and ensure further data normalization through personalized scripts. For instance, one such script is used to collect DICOM date-time fields to guarantee conformance to DICOM standard's expected format. Another script extracts and normalizes dosimetric data from imaging exam data. Each data normalization script collects data stored within DICOM fields and reshapes it to ensure conformance to expected data format from the DICOM standard (commonly by data type conversion with the python standard library). Additionally, these scripts are automatically executed by `Orthanc` on every new DICOM study stored within its database, through a trigger mechanism.

`OpenSearch` - Following data ingestion, curated medical data stored within `Orthanc` are routed to `OpenSearch` in JSON format and in conformance with HTTP protocol (with the HTTP library for python called requests[2]). `OpenSearch`, a fork of Elasticsearch, is an open-source software under Apache 2.0 license offering data indexing, visualisation and exploration functionalities. Curated data elements are stored as documents (lists of data fields containing the relevant DICOM data), which are then clustered in indexes enabling `OpenSearch`'s powerful search engine. For the particular needs of this project, new indexes are generated daily to support expected continuous data ingestion. This index generation is supervised by a specific index pattern, ensuring structural cohesion between indexes.

2.3 Dashboards

Following data ingestion and indexation, dashboards are built within `OpenSearch` to offer flexible and user-friendly interfaces for data exploration, as shown for example in Figure 2. Dashboard offers metrics, graphs and monitoring tools tailored for specific needs. The dashboards can be accessed from any computer, assuming appropriate authentication. It is worth mentioning that dashboards do not report medical data, only aggregated metrics. Several data visualization options are available [3], as well as search and filter tools to drill down in data. Multiple dashboards can be hosted on a single `OpenSearch` instance, each with their distinct dataset. Dashboards are built according to Big Data visualization recommendations [5, 6]

2.4 Data Management

Data management best practices are integrated within PARADIM and throughout the workflow. Chief among them are the FAIR principles[4], as well as a general DataOps philosophy focusing on machine-actionability. These principles and guidelines offer guidance to data scientists and professionals on how data should be managed and integrated. For instance, in our data flows we enforce the preservation of all metadata and we rely on persistent unique identifiers, which is facilitated by the DICOM standard (which also handles the interoperability component of FAIR).

[1] https://paradim.science/

[2] https://requests.readthedocs.io/en/latest/

[3] https://datavizproject.com/

[4] https://www.go-fair.org/fair-principles/

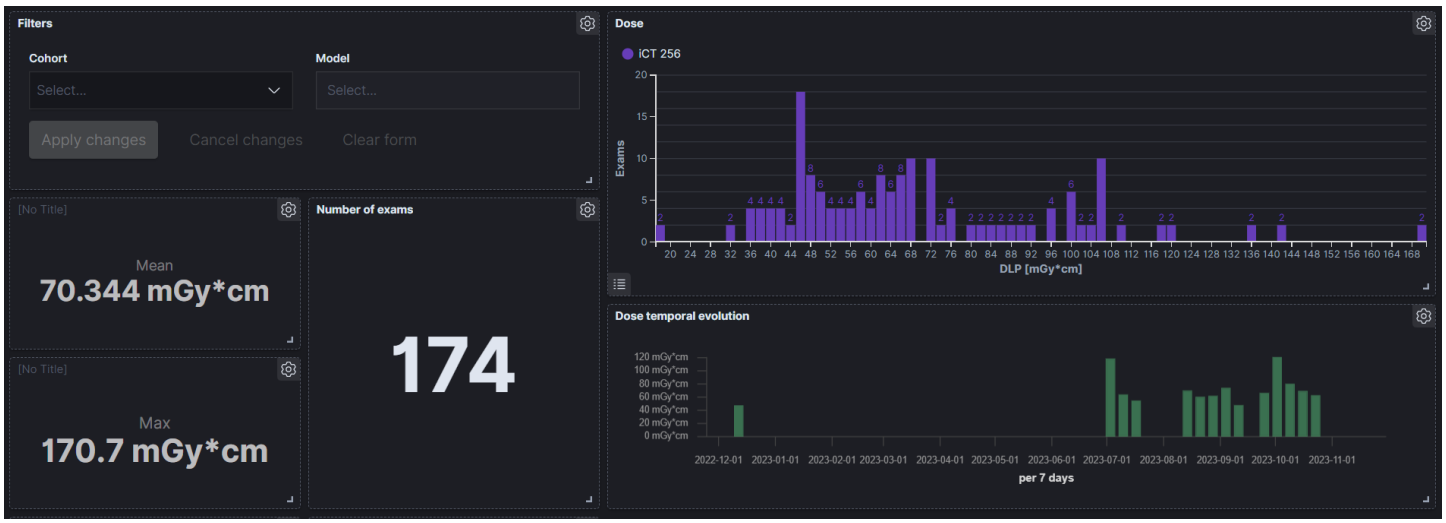

Figure 2: Dashboard example fed by the data pipeline.

3 Results

The robustness of the data pipeline was tested with the open access National Lung Screening Trial database[5]. The platform was able to ingest data at a rate of (5440 ± 145) image series per hour (from the `Karnak` gateway to visibility on dashboards), suggesting that real-time operations are possible for a large institution or even a regional/national setting.
The dashboard example presented in Figure 2 shows some available metrics, including mean and maximum dose, and stacked bar charts (for dose distribution and temporal evolution). Line graphs, pie charts, treemaps, gauge charts and tables can also be used. Additionally, dynamic filters can be integrated into the dashboard, such as the one at the upper left of the dashboard in Figure 2, which allows filtering on device model or clinical cohort. Multiple metrics, such as bar charts, can also be used to drill down to specific values with a single click.
Dashboards provide an overview of radiation usage in CT, leading to opportunities to study clinical practice trends and ease the identification of outliers to improve heath care.

4 Discussion

The proposed infrastructure is well suited to massive datasets and is designed with robustness in mind. PARADIM is deployed in a `kubernetes` environment, where applications are continuously monitored and can be restarted automatically upon failure [7].
The strict adherence to the DICOM standard enforces data normalization and preserves contextual information through rich metadata. Additionally, dosimetric data stored within DICOM radiation dose structured reports (RDSR), such as DLPs, was successfully extracted and ingested by dashboards. However, data kept within private tags is not normalized since their content is not controlled by the DICOM standard.

We have observed that access to rich dosimetric information often depends on the configuration of CT devices. By default, some do not send dosimetric data to the PACS or generate results that cannot be exploited easily (for example dosimetric reports added as images, requiring OCR to extract data, or dosimetry data in non-standard DICOM field). Furhtermore, manufacturers unfortunately offer distinct ways to store dosimetric data despite the availability of the Patient Radiation Dose Structured Report[8, 9]. This heterogeneity is a barrier to large scale dosimetric analyses.

5 Conclusion

A dashboard fed by secure data pipelines was developed to aggregate and visualize dosimetric data from clinical CT studies. The dashboard provides an overview of radiation usage in CT, give rise to opportunities to study clinical practice trends and ease the identification of outliers to improve heath care. Data ingestion is continuous and does not interfere with clinical workflow. Likewise, good data management practices are respected and integrated within the workflow. Moreover, the tool is compliant end-to-end to DICOM standard and uses only open-source, non-proprietary software. Next steps include the development of alerts for events (e.g. high dose, repeated exams) and the integration of data from other centers to generate a personalized report on cumulative dose from CT studies.
Additional work to promote sound data management within health institutions should be considered to maximize the tool's potential. Proper data management include data normalization, compliance to standards (such as DICOM) and richer metadata. Better compliance to the standard by vendors regarding dosimetry appears highly desirable.

[5] https://wiki.cancerimagingarchive.net/display/NLST

References

[1] UNSCEAR. *2008 Report to the General Assembly with Scientific Annexes, Volume I: Sources, Annex A: Medical radiation exposures.* Tech. rep. United Nations, 2008.

[2] A. M. Yi-Sheng Chao Alison Sinclair. "The Canadian Medical Imaging Inventory 2019 - 2020". *Canadian Journal of Health Technologies* 1.1 (2021), p. 215.

[3] H. Canada. "Safety Code 35: Safety Procedures for the Installation, Use and Control of X-ray Equipment in Large Medical Radiological Facilities" (2008), p. 88.

[4] NEMA. *DICOM Attribute Confidentiality Profiles*. URL: https://dicom.nema.org/dicom/2013/output/chtml/part15/chapter_E.html#table_E.1-1.

[5] A. Vellido. "The importance of interpretability and visualization in machine learning for applications in medicine and health care". *Neural Comput & Applic* 32.24 (2020), pp. 18069–18083. DOI: 10.1007/s00521-019-04051-w.

[6] bibinitperiod C. A. A. L. Wang G. Wang. "Big Data and Visualization: Methods, Challenges and Technology Progress". *Digital Technologies* 1.1 (2015). DOI: 10.12691/dt-1-1-7.

[7] A. B. et al. "A scalable, secure, and interoperable platform for deep data-driven health management". *Nat Commun* 12.1 (2021). DOI: 10.1038/s41467-021-26040-1.

[8] G. M. Systems. *Computed Radiography DICOM Conformance Statements*. URL: https://www.gehealthcare.com/products/interoperability/dicom/computed-radiography-dicom-conformance-statements.

[9] S. Healthinners. *DICOM Conformance Statements - SOMATOM Scope*. URL: https://www.siemens-healthineers.com/services/it-standards/dicom-conformance-statements-computed-tomography/somatom-scope.