

 Database

 Credentialed Access

PIFIR: PET-CT Invasive Fungal Infection Reports

Vlada Rozova  , Anna Khanina  , Jeremy Ong  , Ramin Alipour  , Leon Worth  , Monica Slavin  , Karin Thursky  , Karin Verspoor 

Published: Feb. 27, 2025. Version: 1.0.0

When using this resource, please cite: [\(show more options\)](#)

Rozova, V., Khanina, A., Ong, J., Alipour, R., Worth, L., Slavin, M., Thursky, K., & Verspoor, K. (2025). PIFIR: PET-CT Invasive Fungal Infection Reports (version 1.0.0). *PhysioNet*. <https://doi.org/10.13026/d51v-j343>.

Please include the standard citation for PhysioNet: [\(show more options\)](#)

Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R., ... & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation* [Online]. 101 (23), pp. e215–e220.

Abstract

Surveillance of invasive fungal infection (IFI) in clinical settings is a laborious process requiring a detailed review of patient medical history. One of the key sources of clinical information is imaging reports: radiologist-produced free-text reports summarising findings and observations from a scan. Positron emission tomography combined with computed tomography (PET-CT) is a medical imaging modality particularly useful in ruling out IFIs and evaluating the response to anti-fungal therapy. The data was generated to facilitate the development of an automated tool for the detection of IFI. The PET-CT Invasive Fungal Infection Reports (PIFIR) corpus contains 201 de-identified reports annotated by radiologists for terminology suggestive of the presence of IFI. These include IFI-related concepts and certainty cues, as well as relations between them. We release the annotation schema and the original reports along with the corresponding annotation files. We anticipate the PIFIR corpus to be useful in the development and validation of named entity recognition and relation extraction methods with a focus on clinical data. Such methods can be instrumental in processing other types of clinical documentation (clinical notes, nursing notes) with various downstream tasks in mind.

Background

Invasive fungal infections (IFIs) are rare but serious infections most commonly affecting immunocompromised and critically ill patients. IFIs are associated with increased morbidity, mortality, and management costs [1-4]. Routine IFI surveillance is needed in healthcare facilities to allow timely detection of infection outbreaks, identify new and emerging risks for IFI, evaluate infection prevention and prophylaxis interventions, and enable benchmarking between facilities.

The lack of a single diagnostic test means traditional surveillance approaches involve examining multiple sources of clinical information among which imaging reports are a key data element [5]. Positron emission tomography combined with computed tomography (PET-CT) is a medical imaging modality particularly useful in ruling out IFIs and evaluating the response to anti-fungal therapy [6].

Our team aims to design and implement the Invasive Fungal Infection Surveillance (IFIS) tool for automated routine surveillance of IFI, employing a combination of natural language processing (NLP) and machine learning techniques. The PET-CT Invasive Fungal Infection Reports (PIFIR) corpus was created to support the development of an NLP-based classifier to enable the automated processing of imaging reports as a critical component of the IFIS platform.

Methods

Study population and design

PET-CT reports that were undertaken to detect or evaluate the presence of IFI were extracted from patient medical records. The reference list of patients for inclusion was obtained from three studies conducted by clinician researchers at the National Centre for Infections in Cancer, Peter MacCallum Cancer Centre, Melbourne, Australia between March 2010 and July 2020 (Table 1). Reports were incorporated for patients with and without a clinically diagnosed IFI. Only patients receiving treatment for haematological malignancy and/or bone marrow transplant were included.

Table1. Description of clinical studies used to identify relevant PET-CT reports.

Study	Time period	Included reports
Douglas et al [6]	May 2017 - May 2017	PET-CT imaging reports undertaken to detect or evaluate the presence of IFI in patients with clinically diagnosed proven or probable IFI.
Douglas et al [7]	January 2018 - July 2020	PET-CT reports undertaken as part of diagnostic work-up for neutropenic fever for adult patients receiving conditioning chemotherapy for haematopoietic stem-cell transplantation or chemotherapy for acute leukaemia and had persistent or recurrent neutropenic fever.
Tio et al [unpublished]	December 2016 - February 2020	PET-CT imaging reports undertaken to detect or evaluate the presence of IFI in patients with clinically diagnosed proven or probable mould infection.

Report de-identification

Dates were shifted within the text and then removed along with all protected health information, names, initials, and contact details of health professionals and healthcare facilities. The identifying information was replaced with a sequence of "XXXX" of the same length.

Entity and relation annotation

We annotated the data at the level of individual concepts indicating the presence of IFI. The annotation schema was developed iteratively with the intent to capture as much information as possible while maintaining relatively high granularity.

We used brat rapid annotation tool [8] to perform manual annotation. The process involved two radiologists (authors RA and JO) independently annotating a subset of reports followed by a meeting to resolve disparities and update the annotation schema to document consensus. This step was repeated several times until agreement on the structure of the schema, concept definitions, and rules for relations was reached. The annotators then proceeded to apply the final guidelines to the whole corpus of reports. A final consensus meeting was held to resolve any differences in interpretation, mostly around the boundaries of annotated phrases. The final judgement on discrepancies was consolidated into a single set of annotation files.

We adopted the following list of concepts indicating the presence of IFI:

- **Infection or IFI:** clinical query of IFI or any infection in general indicating that its presence is suspected; \recorded in the clinical notes section of the report.
- **Risk factor:** host factors indicating the patient is at a higher risk of developing an IFI.
- **Abnormality:** phrases that refer to a mixture of PET and CT abnormalities suggestive of IFI. Subcategories **Abnormality PET** and **Abnormality CT** were used, where possible, to refer to metabolic activity and structural abnormalities, respectively.
- **Lung:** mentions of the lung where there is an abnormality suggestive of IFI.
- **Sinus:** mentions of sinus where there is an abnormality suggestive of IFI.
- **Other:** location of abnormality suggestive of IFI that is not lung or sinus.
- **Infection/inflammation:** any mention of infection or inflammation in the findings/conclusion. Subcategory **IFI indication** was used specifically for mentions of IFI.

We introduced a relation to capture where in the body each abnormality is seen:

- **location-rel:** links an **Abnormality** (including **Abnormality PET**, **Abnormality CT**) or **Infection/inflammation** (including **IFI indication**) concepts to **Lung**, **Sinus**, or **Other**.

In addition to these, we considered terms expressing certainty and progress:

- **Positive:** affirmative expression.
- **Equivocal:** expression of uncertainty.
- **Negative:** negating expression.
- **Improvement:** changes indicating the condition has improved.
- **Stable:** no changes.
- **Worsening:** changes indicating the condition has worsened.

Both certainty and progress cues were captured only when pertaining to an **Abnormality** or **Infection/inflammation** concepts. They were linked to these target concepts via relations:

- **certainty-rel:** links a certainty cue to an **Abnormality** or **Infection/inflammation** about which it is conveying certainty.
- **progress-rel:** links a progress cue to an **Abnormality** or **Infection/inflammation** which status it is conveying.

Data Description

File overview

- The folder **"reports"** contains the original de-identified reports in the .txt file format.
- The folder **"annotations"** contains brat annotation files in the .ann and .txt file formats.
- The file **"annotation.conf"** is a configuration file for brat that defines concept categories and relations.
- The file **"pifir_metadata.csv"** contains identifiers and a suggestion on how the data should be split for testing and validation.

Each text document has a corresponding annotation file with the same base file name. File names are assigned following the format "report_X", where X is a unique ID assigned to the report. As such, it is not possible to tell from the file names whether two reports belong to the same patient. Instead, patient IDs and scan numbers are provided in the metadata file. Reports belonging to the same patient are numbered chronologically based on the date of the scan (this information is not disclosed).

Annotation files are produced by brat in the brat standoff format (see [9] for a detailed overview; alternatively, files in the .txt format are also available). Information about entities and relations is recorded in the annotation file, one item per line:

- Each entity annotation contains a unique entity ID, its assigned concept category (e.g., ***Risk factor*** or ***Abnormality***), and the span of characters containing the entity mention. For example, "T1 Risk_factor 87 100 AML induction" describes an entity "AML induction" with a unique ID T1, a concept category ***Risk factor***, starting at position 87 and ending at 100.
- Each relation annotation contains a unique relation ID followed by the type of relation (e.g., ***progress-rel***) and its arguments. For example, "R1 progress-rel Arg1:T3 Arg2:T4" describes a relation between entities T3 and T4 with a unique ID R1, a relation type ***progress-rel***, linking its first argument, T3 with its second argument, T4.

Summary statistics

- The total number of reports: 201.
- The total number of patients: 156.
- A follow-up scan is available for 45 patients, for the rest the dataset contains one report per patient.
- The average document length is 1809 characters ranging from 807 to 4464 characters.
- The majority of reports are semi-structured with headers indicating sections such as clinical notes, findings, and conclusion; however, such a layout is not consistent across the corpus.

Usage Notes

The PIFIR corpus was used to develop an NLP-based classifier to detect the evidence of IFI in radiology reports. These results are currently in preparation for submission; the code used to prepare the data and produce the summary statistics above can be found in our GitHub repository [10]. Alternatively, researchers interested in validating their results against ours should refer to "pifir_metadata.csv" for the details on how the data was split in the original study.

It is important to note that the PIFIR corpus was derived from two hospitals within a small geographic area. Collecting more reports positive for IFI was not feasible due to the rarity of the condition. Hence, the number of terms annotated for each category may not be large enough to train (or even fine-tune) a deep-learning NER model. That said, there is a limited number of publicly available clinical text documents, particularly from the Australian context. We thus hope to provide other researchers working on a similar problem with an opportunity to conduct external validation of their methodologies and investigate domain adaptation techniques. We further anticipate the PIFIR dataset to be useful to researchers developing new concept and relation extraction methods with a focus on clinical data. Such methods can be instrumental in processing other types of clinical documentation (radiology reports, clinical notes, nursing notes) with different downstream tasks in mind.

Ethics

Ethics approval granted by Peter MacCallum Cancer Centre (HREC Reference No: HREC/69640/PMCC) and governance approvals from Peter MacCallum Cancer Centre (SSA/69640/PMCC) and Royal Melbourne Hospital (SSA/69640/MH-2020).

Acknowledgements

The authors would like to acknowledge Dr Abby Douglas and Dr ShioYen Tio for providing access to the databases of patients with confirmed infections. This work was supported by the Australian National Health and Medical Research Council(NHMRC) Project Grant APP1156426.

Conflicts of Interest






The authors declare no conflicts of interest.

References

1. Even C, Bastuji-Garin S, Hicheri Y, Pautas C, Botterel F, Maury S, et al. Impact of invasive fungal disease on the chemotherapy schedule and event-free survival in acute leukemia patients who survived fungal disease: a case-control study. *haematologica*. 2010;96(2):337.
2. Pappas PG, Alexander BD, Andes DR, Hadley S, Kauffman CA, Freifeld A, et al. Invasive fungal infections among organ transplant recipients: results of the Transplant-Associated Infection Surveillance Network (TRANSNET). *Clinical Infectious Diseases*. 2010;50(8):1101-11.
3. Neofytos D, Lu K, Hatfield-Seung A, Blackford A, Marr KA, Treadway S, et al. Epidemiology, outcomes, and risk factors of invasive fungal infections in adult patients with acute myelogenous leukemia after induction chemotherapy. *Diagnostic microbiology and infectious disease*. 2013;75(2):144-9.
4. Girmenia C, Raiola AM, Piciocchi A, Algarotti A, Stanzani M, Cudillo L, et al. Incidence and outcome of invasive fungal diseases after allogeneic stem cell transplantation: a prospective study of the Gruppo Italiano Trapianto Midollo Osseo (GITMO). *Biology of Blood and Marrow Transplantation*. 2014;20(6):872-80.
5. Kontoyiannis DP, Marr KA, Park BJ, Alexander BD, Anaissie EJ, Walsh TJ, et al. Prospective surveillance for invasive fungal infections in hematopoietic stem cell transplant recipients, 2001-2006: overview of the Transplant-Associated Infection Surveillance Network (TRANSNET) Database. *Clin Infect Dis*. 2010;50(8):1091-100.
6. Douglas AP, Thursky KA, Worth LJ, Drummond E, Hogg A, Hicks RJ, et al. FDG PET/CT imaging in detecting and guiding management of invasive fungal infections: a retrospective comparison to conventional CT imaging. *Eur J Nucl Med Mol Imaging*. 2019;46(1):166-73.
7. Douglas A, Thursky K, Spelman T, Szer J, Bajel A, Harrison S, et al. [(18)F]FDG-PET-CT compared with CT for persistent or recurrent neutropenic fever in high-risk patients (PIPPIN): a multicentre, open-label, phase 3, randomised, controlled trial. *Lancet Haematol*. 2022;9(8):e573-e84.
8. Stenetorp P, Pyysalo S, Topić G, Ohta T, Ananiadou S, Tsujii Ji, editors. *brat: a Web-based Tool for NLP-Assisted Text Annotation*2012 April; Avignon, France: Association for Computational Linguistics.
9. brat standoff format [Online]. Available from: <http://brat.nlplab.org/standoff> [Accessed Dec 9, 2024]
10. Project GitHub repository [Online]. Available from: <https://github.com/vlada-rozova/pifir> [Accessed Jan 6, 2025]

Contents

Share



Access

Access Policy:
Only credentialed users who sign the DUA can access the files.

License (for files):
[PhysioNet Credentialed Health Data License 1.5.0](#)

Data Use Agreement:
[PhysioNet Credentialed Health Data Use Agreement 1.5.0](#)

Required training:
[CITI Data or Specimens Only Research](#)

Discovery

DOI (version 1.0.0):
<https://doi.org/10.13026/d51v-j343>

DOI (latest version):
<https://doi.org/10.13026/e0cx-1k53>

Topics:

- nlp
- clinical documentation
- information extraction
- invasive fungal infections

Corresponding Author

You must be logged in to view the contact information.

Files

This is a restricted-access resource. To access the files, you must fulfill all of the following requirements:

- be a [credentialed user](#)
- complete required training:
 - [CITI Data or Specimens Only Research](#)
You may submit your training [here](#).
- [sign the data use agreement](#) for the project

PhysioNet is a repository of freely-available medical research data, managed by the MIT Laboratory for Computational Physiology.

Supported by the National Institute of Biomedical Imaging and Bioengineering (NIBIB) under NIH grant number R01EB030362.

For more accessibility options, see the [MIT Accessibility Page](#).

[Back to top](#)