Building a reference multimedia database for interstitial lung diseases

Adrien Depeursinge [a,b,*], Alejandro Vargas [b], Alexandra Platon [c], Antoine Geissbuhler [b],
Pierre-Alexandre Poletti [c], Henning Müller [a,b]

[a] *University of Applied Sciences Western Switzerland (HES-SO), TechnoArk 3, CH-3960 Sierre, Switzerland*
[b] *University and University Hospitals of Geneva, Service of Medical Informatics, 4, rue Gabrielle-Perret-Gentil, CH-1211 Geneva 14, Switzerland*
[c] *University and University Hospitals of Geneva, Service of Emergency Radiology, 4, rue Gabrielle-Perret-Gentil, CH-1211 Geneva 14, Switzerland*

ARTICLE INFO

ABSTRACT

This paper describes the methodology used to create a multimedia collection of cases with interstitial lung diseases (ILDs) at the University Hospitals of Geneva. The dataset contains high-resolution computed tomography (HRCT) image series with three-dimensional annotated regions of pathological lung tissue along with clinical parameters from patients with pathologically proven diagnoses of ILDs. The motivations for this work is to palliate the lack of publicly available collections of ILD cases to serve as a basis for the development and evaluation of image-based computerized diagnostic aid. After 38 months of data collection, the library contains 128 patients affected with one of the 13 histological diagnoses of ILDs, 108 image series with more than 41 l of annotated lung tissue patterns as well as a comprehensive set of 99 clinical parameters related to ILDs. The database is available for research on request and after signature of a license agreement.

1. Introduction

In order to understand the underlying mechanisms of groups of diseases, the first requirement is to collect a sufficient number of cases that are representative of the various realizations of the studied diseases. Moreover, the quality of the acquired data is a key for carrying out non-biased evaluations of computer-aided diagnosis (CAD) systems. With the aim to enhance the disease therapy in clinical routine, the library of cases has to be as representative as possible of the hospital's population, meaning that the cases have to be chosen randomly among the entire population if possible.

These requirements are of high importance when building image-based computerized diagnostic aid based on medical image processing [1,2]. A high-quality multimedia collection of cases containing annotated image series and associated clinical parameters is required to ensure the success of a CAD system at the time it will be integrated into clinical routine [3].

On the one hand, the database constitutes a basis for developing computerized tools [4,5] such as automatic detection of abnormal pulmonary tissue types in high-resolution computed tomography (HRCT) images and retrieval of similar cases [6,7]. Detailed ground truth and a large number of cases allow to reliably evaluate and compare medical image processing algorithms for the defined tasks

(i.e., benchmarks). Popular datasets such as Lena, Brodatz [8] or Iris[1] allow to qualitatively and quantitatively evaluate a large number of basic methods in image processing and machine learning and thereby established a *de facto* reference dataset.The popularity of these datasets is partly due to the fact that they reflect real-life challenges thus offering more credibility of the obtained results when compared to artificial datasets. It is also important that with such publicly available data sets results can be reproduced by other researchers and be compared to the state of the art, an important requirement in science. On the other hand, such databases also create opportunities for specialized studies and teaching. The cases with confirmed diagnoses constitute a knowledge base that can be used as diagnostic aid [9]. For instance, advanced browsing enabled by content-based image retrieval (CBIR) or multimodal case-based retrieval from large databases of cases with confirmed diagnoses can be highly valuable for radiologists with little experience in the domain [6,10,11]. In summary, a high-quality multimedia library of cases is valuable for:

- teaching,
- specialized descriptive studies,
- training and testing pattern recognition techniques,
- retrieving similar cases as diagnostic aid,

---

* Corresponding author at: University of Applied Sciences Western Switzerland (HES-SO), TechnoArk 3, CH-3960 Sierre, Switzerland. Tel.: +41 27 606 9033.
E-mail address: adrien.depeursinge@hevs.ch (A. Depeursinge).

[1] http://archive.ics.uci.edu/ml/datasets/Iris/, as of 24 May 2011.

- comparative performance analysis of medical image processing methods (i.e., benchmarks [12]).

However, the construction of a high-quality multimedia collection of cases is extremely time-consuming and expensive. It is often a bottleneck in studies on image-based CAD systems. The identification of the relevant cases, the consultation of the electronic health record (EHR) and picture archiving and communication systems (PACS) to gather the clinical parameters and the image series, the data entry as well as the database infrastructure and maintenance involve a large amount of work with a wide range of skills from medical knowledge to information technology (IT) expertise. To assess the high-quality of the data, several researchers and physicians have to be involved in the case selection process and the delineation of regions of interest (ROIs) to cope with the inter- and intra-observer variabilities, the latter being particularly important in radiology [13,14]. The agreement of the ethics committee has to be obtained before starting any investigations. The latter constitutes the required justification to access content of the EHR.

Depending on the studied diseases, cases are sometimes rare and are encountered casually even in large university hospitals. Incidentally, these diseases are the ones requiring reference databases to palliate the lack of experience due to their sparsity. Efforts from the European Union (EU) research programme came up with information technology (IT) infrastructures to cope with the difficulty of collecting rare cases with aneurysms in the AneurIST project[2] [15]. Multimedia data from six clinical centers within Europe was gathered. Anonymization of the patient data as well as images [16] is required as soon as the data leaves the medical institution [17].

A major observation when studying the state-of-the-art of texture-based CADs for lung tissue analysis in thin-section computed tomography (CT) is the lack of statistical significance of the measured performance as the CAD systems are most often evaluated on a small number of cases [18]. In this paper, the steps for building a multimedia database of cases with interstitial lung diseases (ILDs) and the current content of the created database are detailed. This database was built in the context of the Talisman[3] project [19].

1.1. Interstitial lung diseases

ILDs can be characterized by the gradual alteration of the lung parenchyma leading to breathing dysfunction. They regroup more than 150 histological diagnoses associated with disorders of the lung parenchyma [20]. The factors and mechanisms of the disease processes vary from one disease to another and the exact cause of many ILDs is still unknown [20]. Physical examination of a patient affected by ILD is frequently abnormal but with unspecific findings. The diagnosis of these pathologies is established based on the complete history of the patient, a physical examination, laboratory tests, pulmonary function testing (PFT) as well as visual findings on chest X-ray.

Images play an important role for confirming the diagnosis and patients may not require surgical lung biopsy when the clinical and radiographic impression is consistent with a safe diagnosis [21]. The first imaging examination used is the chest radiograph because of its low cost and weak radiation exposure. It also provides a quick overview of the entire chest. However, chest radiographs are normal in more than 10% of the patients with some forms of ILD and

Table 1
The HRCT scanning protocol.

Slice thickness	1–2 mm
Spacing between slices	10–15 mm
Scan time	1–2 s
Lung shape	Inspirium
Contrast agent	None
Axial pixel matrix	512×512
x, y spacing	0.4–1 mm

can provide a confident diagnosis in only 23% of the cases with lung diseases in general [22]. When the synthesis of this information arouses suspicions toward an ILD, HRCT imaging of the chest is often required to acquire a rapid and accurate visual assessment of the lung tissue. Indeed, the three-dimensional form of HRCT data avoids superposition of organs and provides an accurate assessment of the patterns and distribution of the lung tissue with a submillimetric resolution. It quickly became the gold standard imaging protocol for the diagnosis of diffuse pulmonary parenchymal diseases.

The most common histological diagnoses of ILDs according to [20,23] are detailed in Fig. 1. The associated lung tissue patterns in HRCT are listed in Table 2.

1.1.1. High-resolution computed tomography of the lung

In 1972, the first commercial CT scanner created image series with relatively low resolution of an 80×80 pixel matrix in each axial slice. The numerical value of each pixel is related to the X-ray attenuation and is expressed in Hounsfield Units (HU). Modern scanners with multiple detectors use a helical scanning mode and can provide a 3D array of isotropic voxels with a submillimetric resolution. This protocol is called multidetector computed tomography (MDCT). The main drawback of this protocol is the high amount of radiation dose the patient is exposed to. To assess the visual appearance of healthy and pathological lung tissue, a submillimetric resolution is required but some areas can be skipped between the thin sections to limit the radiation exposure. This protocol is called HRCT and is the gold standard imaging protocol for diagnosing ILDs [24]. The technical specifications of the HRCT protocol are listed in Table 1.

HRCT is also more appropriate than magnetic resonance imaging (MRI) to assess the visual appearance of the lung tissue. Indeed, MRI is only sensitive to inflammatory changes of the pulmonary parenchyma as other tissue types have a low density of protons [25]. A comparison between HRCT and MRI is shown in Fig. 2.

1.1.2. Lung tissue patterns associated with ILDs in HRCT

The appearance and quantification of the types of lung tissue patterns in HRCT are very informative for establishing the differential diagnosis of an ILD. Table 2 lists 13 common histological diagnoses of ILDs, the associated HRCT findings as well as the region of the lungs where the disease is predominant. The visual aspects of the most common lung tissue patterns are depicted in Fig. 3. The taxonomy used to describe them often relates to texture properties. The term *fibrosis* is used in this work to describe all HRCT findings that are associated with the histological diagnosis "pulmonary fibrosis" and includes *reticulation, traction bronchiectasis, architecture distortion* and *honeycombing* [26]. As observed in Table 2, *ground glass* patterns are encountered in most of the ILDs and are thus non-specific. Therefore, the clinical context and other HRCT findings are required to orient the diagnosis.
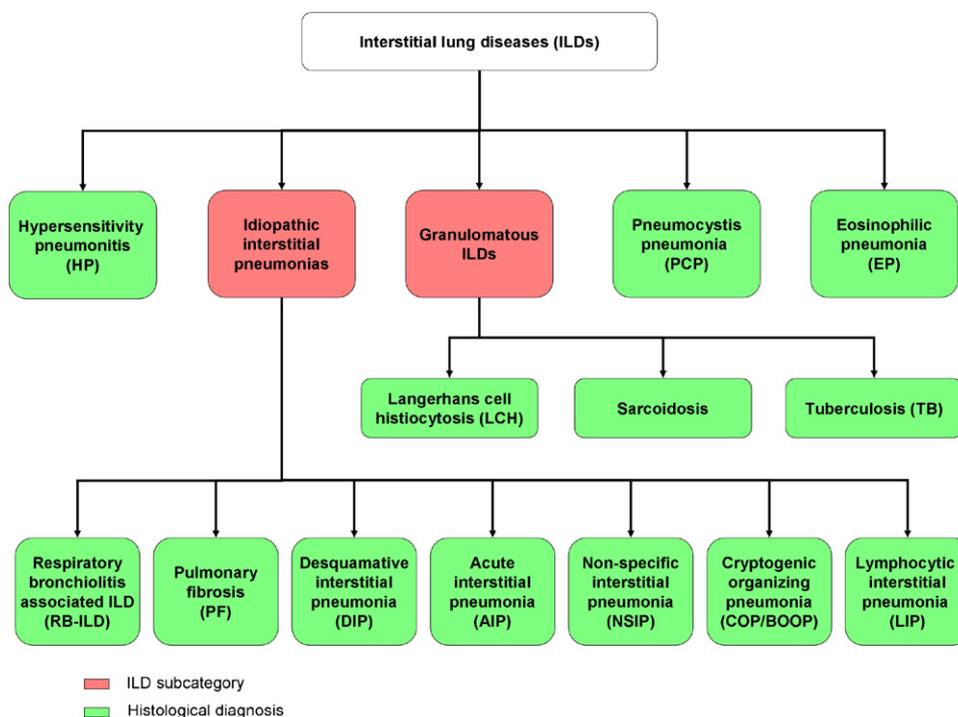
Fig. 1. The most common histological diagnoses of ILDs according to [20,23].

(a) HRCT   (b) 3–Tesla MRI
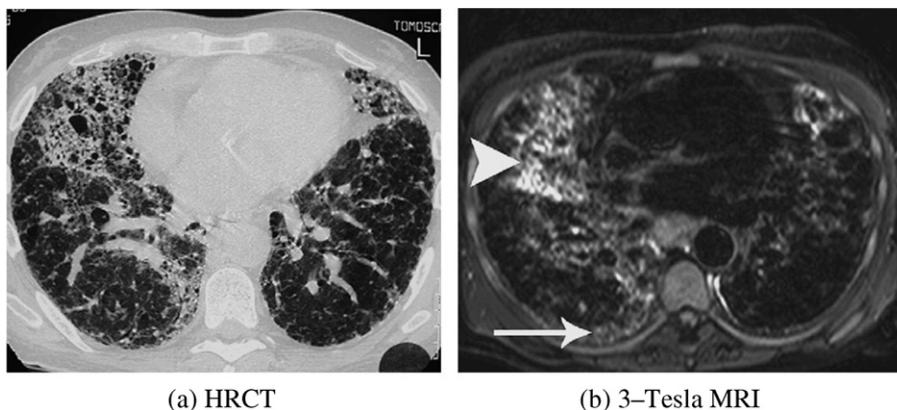
Fig. 2. A comparison of HRCT and MRI for the visual assessment of the lung tissue (*Source*: [25]). In the MRI image (b), the signal of the infected lung-area (arrowhead) is 120% higher than the fibrous tissue area (arrow).

Table 2
13 common histological diagnoses of ILDs and associated HRCT findings.

Histological diagnosis	HRCT lung tissue patterns	Predominance
Hypersensitivity pneumonitis (HP)	*Ground glass, emphysema, fibrosis*	Diffuse
Pneumocystis pneumonia (PCP)	*Ground glass, crazy-paving, cysts, pneumothorax*	Central, perihilar
Eosinophilic pneumonia (EP)	*Ground glass, consolidation, crazy-paving*	Peripheral, apex
Langerhans cell histiocytosis (LCH)	*Cysts, Ground glass, micronodules, reticulation*	Apex
Sarcoidosis	*Micronodules, consolidation, macronodules, ground glass, fibrosis* (end-stage)	Peribronchovascular, subpleural, peripheral
Tuberculosis (TB)	*Micronodules* (miliary), *tree-in-bud, consolidation*	Diffuse
Respiratory bronchiolitis associated ILD (RB-ILD)	*Ground glass, emphysema*	Diffuse, centrilobular
Pulmonary fibrosis (PF)	*Fibrosis, bronchiectasis, ground glass*	Peripheral, subpleural, basal, posterior
Desquamative interstitial pneumonia (DIP)	*Ground glass, emphysema, fibrosis* (uncommon)	Subpleural, basal
Acute interstitial pneumonia (AIP)	*Ground glass, consolidation*	Basal, diffuse
Non-specific interstitial pneumonia (NSIP)	*Ground glass, consolidation, reticulation, fibrosis* (uncommon)	Peripheral, basal
Cryptogenic organizing pneumonia (COP /BOOP[a])	*Ground glass, consolidation* (patchy), *macronodules, macronodules, bronchial wall thickening, crazy-paving*	Peribronchovascular, subpleural
Lymphocytic interstitial pneumonia (LIP)	*Ground glass, micronodules*	Peribronchovascular, subpleural

[a] Bronchiolitis obliterans organizing pneumonia (BOOP) was formerly used and replaced by COP.

(a) *healthy* (b) *emphysema*

(c) *ground glass* (d) *fibrosis*

(e) *micronodules* (f) *consolidation*
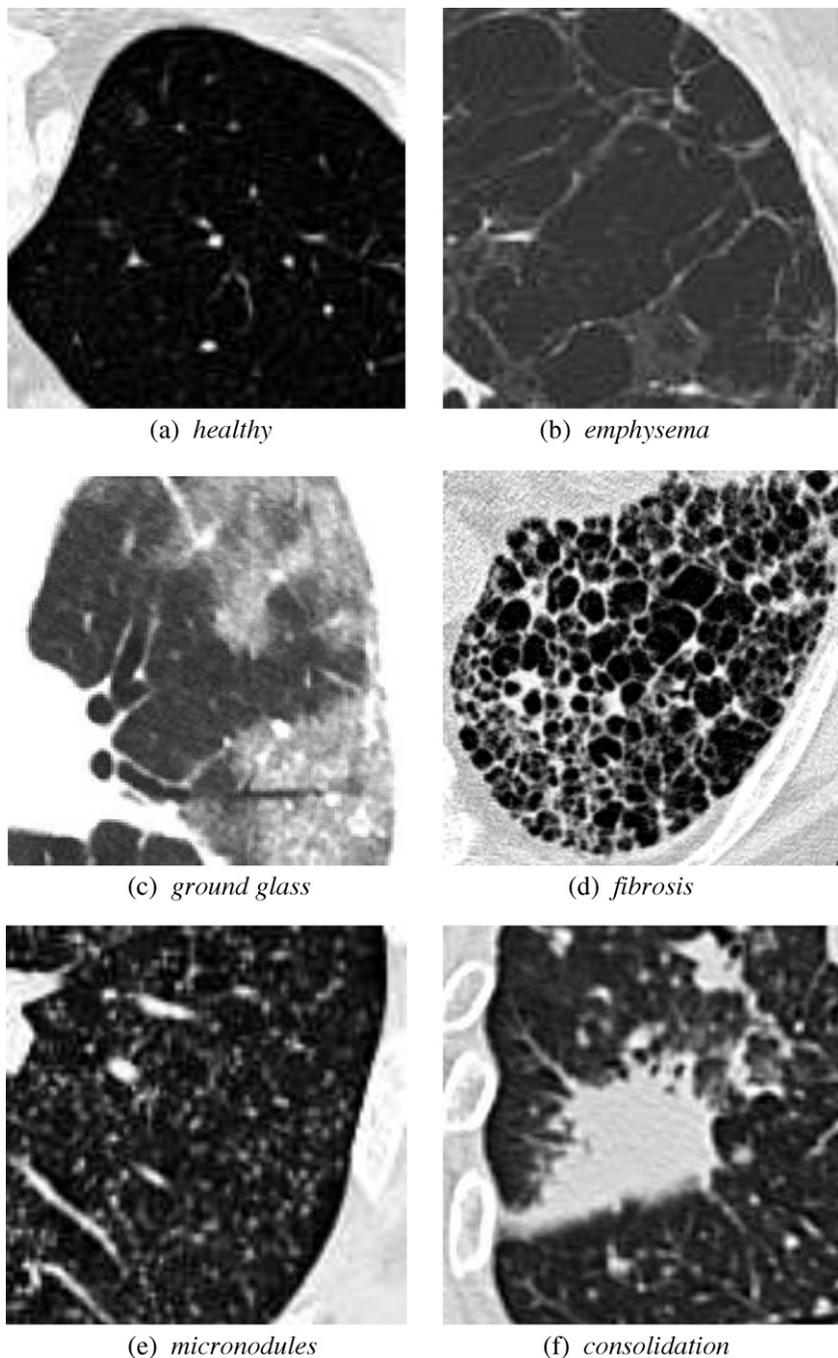
Fig. 3. Visual aspects of the most common lung tissue patterns in HRCT of patients with ILDs. (a) *Healthy*; (b) *emphysema*; (c) *ground glass*; (d) *fibrosis*; (e) *micronodules*; (f) *consolidation*.

1.1.3. Interpretation of HRCT image series

Interpreting HRCT images of the chest represents a challenge even for trained chest radiologists and lung specialists [22,23]. The three-dimensional form requires significant reading time, effort, and experience for a correct interpretation [27]. Most often, the interpretation process is carried out by comparing a case with similar images in textbooks such as [23] or with similar cases in personal image collections, which are most often organized by pathology. To do so, the radiologists must have a guess of the suspected disease present in the image and may miss the true pathology shown. In certain medical services (e.g., emergency radiology service), radiologists have recourse to a large diversity of imaging modalities such as conventional projection radiography, CT, MRI, functional imaging (functional MRI (fMRI), positron emission tomography (PET)), and ultrasound applied to different organs such as the brain, colon, breast, chest, liver, kidney and the vascular and skeletal systems. They have to provide the first radiological report with ideas on the diagnosis quickly. This may result in errors by omission or confusion of diverse pathologic lung tissue types [28]. Moreover, the context is fundamental for correct interpretation: healthy tissue, for example, may have different visual aspects depending on the age or the smoking history of the patient and *ground glass* findings are non-specific without complementary clinical parameters [29].

1.2. Existing databases of CT imaging of the lung

Efforts for building a resource for the lung imaging research community are detailed in [30,31]. In order to test and develop lung CADs from reliable datasets for the detection of lung nodules on CT scans, the lung imaging database consortium (LIDC) was constituted of five academic institutions from across the United States. The database includes healthy and pathologic CT images with annotated nodules and primary clinical data of the patient. Expert radiologists from the five institutions agreed on the definition of nodules and the criteria for inclusion in the database. The nodules are categorized into three classes based on their diameters: "nodule ≥3 mm", "nodule <3 mm" and "non-nodule ≥3 mm". The annotations of up to five radiologists are available. After a blinded session of annotations, the radiologists had access to the annotations of their colleagues and had the possibility to retrospectively modify their own. The number of nodules on which all four radiologists agreed was 33.8% for the blinded session and 45.8% after review [14]. The database is publicly available and can be downloaded online from the national biomedical image archive[4] (NBIA). Unfortunately, the LIDC database does not contain ILD cases as it only focused on nodules in CT imaging. A small database of annotated nodules in CT imaging of the lung is also publicly available with the purpose of comparing CAD performance described in [32]. This database can be downloaded online.[5]

Similar efforts are found at the National Heart, Lung, and Blood Institute (NHLBI) but focusing on lung tissue with the creation of the lung tissue research consortium (LTRC[6]) [33]. The goal of LTRC is to improve the management of diffuse lung diseases through a better understanding of the biology of chronic obstructive pulmonary diseases (COPDs) and fibrotic ILDs including idiopathic pulmonary fibrosis (IPF). Control cases were also enrolled. The database aims at creating an open data set containing histological, clinical and radiological data. The LTRC began recruitment in February 2005 and are ahead of their goal of collecting 1600 subjects with a total of 1844 enrolled as of September 30, 2008. The lung tissue patterns are described in a structured report. However, based on [33] no regions of interest were delineated in the image series to serve as ground truth for the evaluation of computerized categorization of the lung tissue or as teaching examples. Free access to the HRCT image series and associated metadata is possible after obtaining the approval of the LTRC data coordinating center based on the submission of a concept sheet describing the aim of the study using a standardized format. Based on the LTRC data, computerized quantification of the disease patterns was proposed as measures of the extent of pulmonary disease in [34].

A web-based teleradiology framework for acquiring cases of diffuse lung diseases with annotated regions is proposed in [35]. The Learning Medical Image Knowledge (LMIK) collaborative platform provides tools to the clinicians to delineate ROIs in HRCT imaging of the chest. The cases are then anonymized and stored in a central database that can be queried by authorized researchers for CAD evaluation and by radiologists for teaching. Unfortunately, no public access to the case repository is mentioned and no recent report on the LMIK activities has been found since 2003.

To the best of our knowledge, besides these efforts, no large dataset with annotated image regions is available to be used as ground truth for the evaluation and comparison of computerized categorization of lung tissue in HRCT.

2. Methodology

In this section, the scope of the database as well as the various steps of data collection is detailed. The methods described are the results of iterative refinements of the process during the 38 months of data collection.

2.1. Scope of the database

Before any collection of cases, the scope of the database was defined in order to obtain a consistent set of cases with the aim of building computerized diagnosis aid for ILDs. In this section, the selection of the histological diagnoses to be included in the database and the associated clinical parameters are detailed.

2.1.1. Selection of the histological diagnoses

In collaboration with the Service of Emergency Radiology and the Service of Pneumology of the University Hospitals of Geneva (HUG), 15 histological diagnoses that are known as the most frequent causes of lung parenchymal disorders were selected [36]. The objective was to retrospectively analyze at least 150 cases representative of the 15 most frequent ILDs from the EHR at the HUG during the four years of the Talisman project. For two diagnoses, no pure case was found during the project resulting in the 13 diagnoses listed in Fig. 1. Although healthy cases that underwent an HRCT exam are rare, the latter were included as often as possible to serve as control cases.

2.1.2. Selection of the clinical parameters

Based on each pathology, the most discriminative clinical parameters for the establishment of the differential diagnosis were kept. This selection process was carried out based on the domain-specific literature [20,23] along with knowledge bases of computer-based diagnosis decision support systems [9]. Discussions and remarks from lung specialists, radiologists and the medical informatics research group (Service of Medical Informatics, HUG) allowed an iterative review of the selected parameters as well as standardized units and data formats to be used. The parameters that were not available from the EHR (such as race) were removed. After several modifications of the list based on the availability of the parameters in the EHR, 159 fields were used to characterize the subgroup of ILDs. A HTML (hypertext markup language) form was used to capture the clinical parameters.[7] The terminology used in the HTML form is in accordance with MeSH[8] and SNOMED-CT[9] medical terminology references. As often as possible, pull-down menus were used for textual data to favor homogeneity of the data required for further computerized analysis. Units for laboratory tests and other numerical data were chosen depending on the formats used in the electronic patient record at the HUG.

2.2. Data collection

The selection process of cases, the annotation of the images and data entry are the result of regular discussions involving the radiologists, the research physicians and the computer scientists during the four years of the project. The several steps of the construction

---

Fig. 4. A typical example of non-accurately delineated ROIs. The ROI contains as much healthy tissue as pathological, introducing noise in the training data (*Source*: [37]).
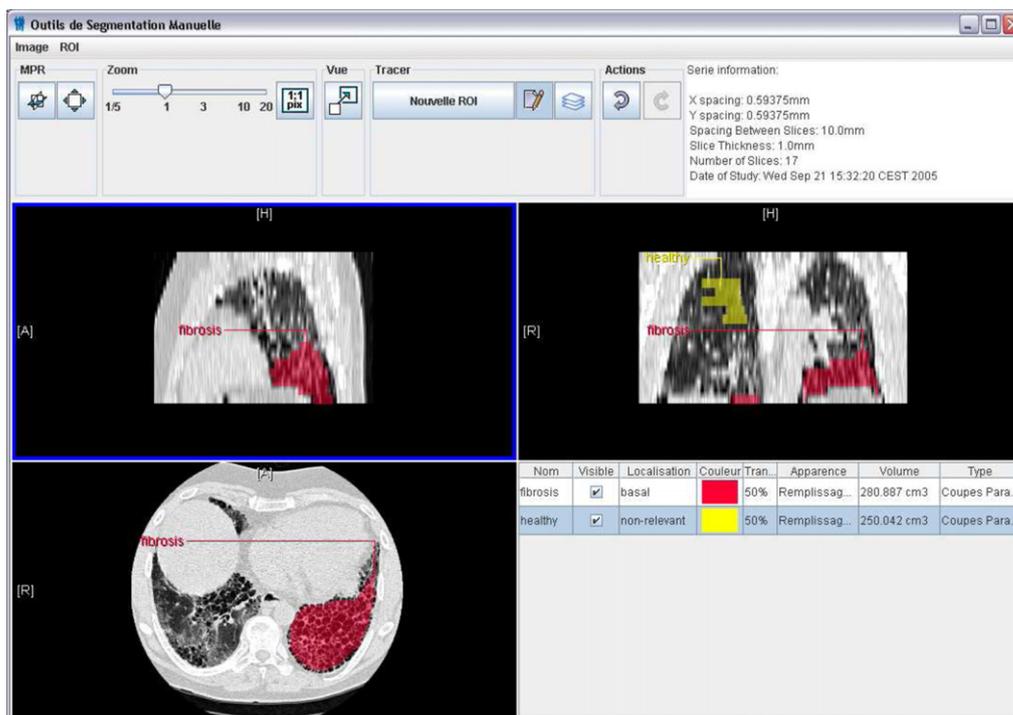
Fig. 5. A screenshot of the graphical tool for the annotation of image regions.

of the multimedia dataset starting from the selection of the cases to the data entry are detailed in this section.

2.2.1. Selection of the cases at the University Hospitals of Geneva

A raw list of 1266 patients that underwent a thorax CT within a stay in the pneumology service between 2003 and 2008 was extracted from the data repository of the EHR by the helpdesk at the medical informatics service. Only cases with HRCT (without contrast agent, 1 mm slice thickness) were retrospectively analyzed. Cases from paediatrics were left aside. The diagnosis of each of the remaining cases was retraced in the EHR based on clinical history, reports and clinical examinations.

At first, the discharge summary and the pneumological consultation report[10] were revised to decide whether the cases can potentially contain an ILD. When the reports of the CT scans were consistent with the clinical reports, the clinical history, and the laboratory studies (pathology, pulmonary function testing, hema-

tologic tests) the cases were nominated as candidates. Cases of which the histological diagnosis is confirmed to be one of the 13 listed in Fig. 1 by at least one of the pathological exams (biopsy, bronchoalveolar lavage (BAL)) were selected for inclusion. When the radiographic impression was consistent with the verified diagnosis, the case was retained for the annotation sessions with the radiologists. These cases were subsequently studied from the radiological point of view during regular meetings with two attending radiologists. A very selective process retaining only cases with high confidence was required in order to gain time at the annotation sessions and concentrate on the radiological aspects only.

The time necessary to decide whether the case must be kept and annotated varied from 15 min to 2 h, with an approximate average time of one hour per case. The selection process at the early months of the project was longer as the methodology was still not well established. It was usually quicker to reject a case than to decide to keep it as ruling out an ILD diagnosis or discovering a large number of co-morbidities can be quick.

Up to now, more than 700 cases were revised and 128 were stored in the database (approximately 18% of the cases had an ILD). Occasionally, patients that had a confirmed diagnosis of ILD

---

10 based on their availability.
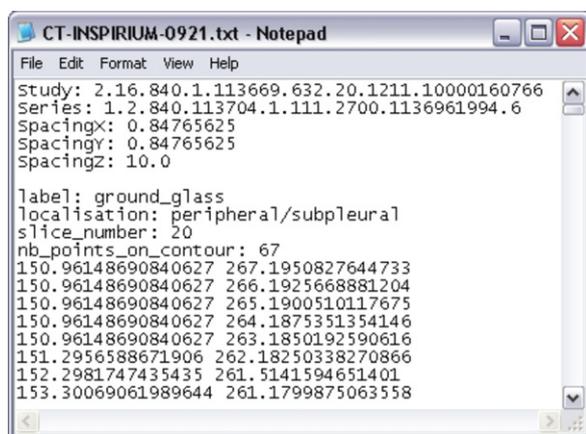
Fig. 6. An example of the text-based file format to store the ROIs in the database.

but from which the HRCT examination was not corresponding to the disease episode were kept. Consequently, 108 cases have an annotated HRCT image series and 20 have no images associated.

2.2.2. Annotations

The purpose of HRCT annotation is to establish the ground truth for lung tissue classification as well as to show examples of HRCT findings related to a studied disease for teaching purposes. The need for high-quality annotations was highlighted in [37]. Indeed, since the annotations are intended for computerized pattern recognition, the ROIs have to delineate pathologic patterns very precisely to avoid the introduction of noise in the training data (see Fig. 4).

The possibility to visualize and delineate three-dimensional ROIs in the entire HRCT volume and to set the window level used for displaying the 16-bit DICOM (Digital Imaging and COmmunications in Medicine) image series on a computer screen was required for annotation. These specifications were fulfilled by adapting an existing graphical software originally developed for delineating hepatic tumors in CT scans at the HUG (see Fig. 5). The radiologist opens an entire DICOM series and then draws precise ROIs in any layer of the CT volume in the axial view. 3D ROIs can be drawn by linearly interpolating the regions between 2D ROIs delineated in non-contiguous axial slices. Sagittal and coronal views are only available for visualization. Depending on the spacing between slices used, anisotropy in the vertical direction prevents delineating ROIs in sagittal and coronal views. No exhaustive annotation of every patterns contained in an entire HRCT scan is performed, only patterns that are related to the disease of the patient are delineated. A very simple text file format was developed to save or load ROIs. The coordinates of the points belonging to the contour of polygons demarcating the ROIs in 2D slices are stored. An example is depicted in Fig. 6. The ROI files were also translated to the binary file format *.seg used in YaDiV[11] to facilitate the handling of 3D ROIs. These files are based on "BitCubes", where a 1 represents a voxel that does belong to this segment and 0 otherwise. The bits are stored in an integer array where width of the slice $xSize$ modulo 32 forms a row and the total number of integers is $N_{rows} \times ySize \times zSize$. It is a trade-off between memory efficiency and operation optimization.

Patterns used to describe lung parenchymal disorders are not standardized among radiology communities. A detailed description of common patterns is given in [22]. The terminology used in this work was mainly derived from [20,23] and a list of the annotated patterns can be found in Table 5. The first set of lung tissue types

Table 3
Terminology used to describe the locations of HRCT findings.

Name	Description
Apical	Upper region
Basal	Lower region
Diffuse	Uniformly distributed
Perihilar	Middle region, around the mediastinium
Peripheral/subpleural	Lung periphery/under the pleural membrane surrounding the lobes
Non-relevant	Used when the pattern has no prevailing location

associated with the 15 initially selected diagnoses was defined with the radiologists and was adapted to the annotation requirements through the annotation sessions. Localization of the parenchymal disorders is relevant for several diseases. Thereby, the localizations of the ROIs are stored along with pattern labels in the ROI files. Table 3 lists the localizations used.

Series with several co-morbidities, or with blur caused by breathing or movements of the patient or containing artifacts were not selected for annotations. Some images taken with contrast agent were also annotated and stored. When possible, healthy tissue was delineated in the studied series to provide a wide range of the aspects of normal lung parenchyma.

The average time for annotating one case was approximately 75 man-minutes. 30 min were necessary for the two radiologists (i.e., 60 man-minutes) to interpret the HRCT image series and to draw coarse annotations highlighting the important events in the series. Another 10–15 min were then required to refine the annotations and obtain accurate delineations of the lung tissue patterns as well as to capture them in the database using a Java-based tool, which was carried by one radiologist.

2.2.3. Data entry

The content of the EHR is systematically analyzed to fill the content of the HTML form consistently. When multiple instances of clinical parameters (e.g., laboratory data) were available in the EHR within an interval of two weeks around the date of the HRCT image series, the instance as close as possible to the HRCT examination was retained. For each HRCT image series stored in the database, the entire set of clinical parameters was filled and cases that have several relevant image series have several instances in the database as soon as the HRCT examinations were not corresponding to the same disease episode or if they were separated by more than two weeks.

The discharge summary and any free-text documents that contained evidence of the diagnosis were anonymized and stored in the database. The EHR at the HUG contains computer tools for the automatic anonymization of the documents. No confidential data is stored in the extracted database, except the patient and stay numbers, which both require authorized access to the EHR to retrieve the identity of the corresponding patient. When transferring the database externally these numbers were replaced by consecutive numbers.

Three medical doctors (MD) were successively responsible for the selection of the cases and the data entry. The final protocol described in Section 2.2.1 was developed and refined over time by the three MDs. The time necessary to capture a case was on average of 75 min, varying from 40 min to more than two hours mostly depending on the diseases. Patients with Sarcoidosis were quickly completed whereas IPF, HP and AIP were requiring more efforts to retrace the history of the patients and to gather all associated parameters. An adaptation period was necessary for the MDs to get used to the EHR at the HUG, as well as to gain experience with the various ILD diagnoses.

---

[11] YaDiV: Yet Another DIcom Viewer, http://www.welfenlab.de/en/research/projects/yadiv/, as of 24 May 2011.

Histological diagnoses
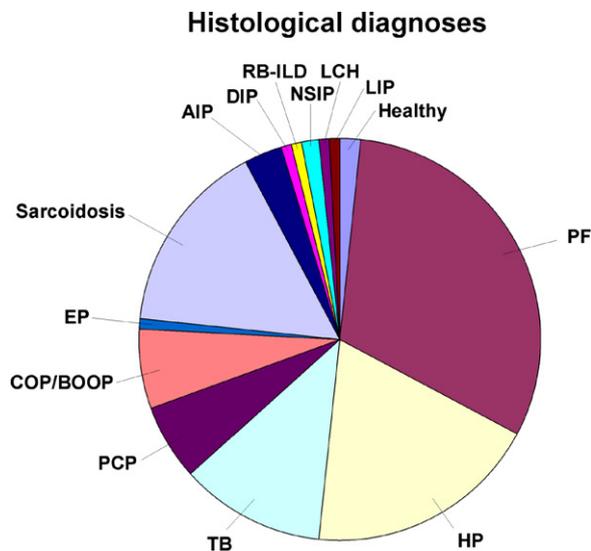
Fig. 7. Distribution of the histological diagnoses. The database contains 128 cases of which 84.37% (108) have an annotated image series.
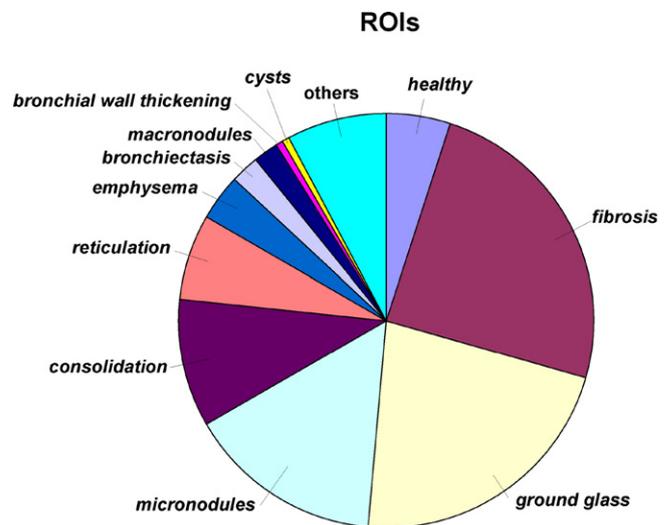
ROIs

Fig. 8. Distribution of the various lung tissue patterns in terms of hand-drawn ROIs. The total number of ROIs is 1946.

Pattern volumes

Fig. 9. Distribution of the various lung tissue patterns in terms of volumes of annotated tissue. The total volume is 41.65 l.

3. Results

In this section, the content of the database resulting from 38 months of data entry is quantitatively described and analyzed. The first case was captured on the 21st of July 2006 and the last on the 6th of November 2009. The process of data quality control and correction is described in Section 3.2. Several examples of data exploitation are listed in Section 3.4.

3.1. Numbers and statistics

128 cases are currently captured in the database. Among them, 108 have an annotated HRCT image series corresponding to the correct period of care. 68.8% (88 cases) underwent a biopsy, 64.8% (83 cases) present a BAL. Of the 128 cases, 21.1% (27 cases) have neither a biopsy nor a BAL but had a specific test confirming the diagnosis (e.g., tuberculin skin test for TB, and Kveim test for Sarcoidosis). The distribution of the diagnoses is detailed in Table 4 and Fig. 7. 1946 ROIs were delineated in 108 HRCT image series resulting in a total volume of 41.65 l of annotated tissue. The distributions of the number of ROIs and the corresponding volumes of the lung tissue types are detailed in Table 5 and Figs. 8 and 9.

3.2. Quality control

A retrospective control of the diagnosis of each case was carried out to ensure the consistency of the database. Since the global

Table 4
Distribution, mean age and gender statistics of the diagnoses.

Histological diagnosis	Patients	Image series	Age (mean ± std)	Female (%)
Healthy	2	2	63.5 ± 9.5	100
Pulmonary fibrosis (PF)	40	39	71.4 ± 13.4	50
Hypersensitivity pneumonitis (HP)	24	19	65.3 ± 17.1	80.3
Tuberculosis (TB)	15	12	41.1 ± 17.6	40
Pneumocystis pneumonia (PCP)	8	4	59.6 ± 20.4	12.5
Cryptogenic organizing pneumonia (COP/BOOP)	8	3	45 ± 24.6	50
Eosinophilic pneumonia (EP)	1	1	33	100
Sarcoidosis	20	18	48.5 ± 17.3	30
Acute interstitial pneumonia (AIP)	4	4	65.3 ± 5.5	50
Desquamative interstitial pneumonia (DIP)	1	1	46	100
Respiratory bronchiolitis associated ILD (RB-ILD)	1	1	54	100
Non-specific interstitial pneumonia (NSIP)	2	2	61.5 ± 12.5	50
Langerhans cell histiocytosis (LCH)	1	1	24	0
Lymphocytic interstitial pneumonia (LIP)	1	1	32	0
Total	128	108	59 ± 20.2	36.7

Table 5
Distribution of the lung tissue patterns. Note that the total number of image series is not equal to the sum of the number of series per pattern as each image series may contain several lung tissue types.

Label	Volume (l)	Image series	ROIs	Mean volume per ROI (10^{-2} l)
Healthy	5.12	7	100	5.12
Fibrosis	8.45	38	473	1.79
Ground glass	4.91	37	427	1.15
Micronodules	16.06	16	297	5.41
Consolidation	0.69	14	196	0.35
Reticulation	1.88	10	131	1.43
Emphysema	1.15	5	66	1.75
Bronchiectasis	0.08	8	44	0.18
Macronodules	0.18	7	37	0.49
Bronchial wall thickening	0.01	1	15	0.08
Cysts	0.15	3	11	1.40
Others	2.95	14	149	–
Total	41.65	108	1946	–

Table 6
Estimations of the person-hours spent for the various tasks of the database construction.

Task	Time computation	Time spent (person-minutes)	Time spent (person-hours)
Cases captured with annotated image series	108 cases × 210 min	22,680	378
Cases captured without image series	20 cases × 135 min	2700	45
Cases erased after quality control	21 cases × 210 min	4410	73.5
Correction of clinical parameters after quality control	13 cases × 75 min	975	16.25
Rejected cases	(700 − 149) cases × 60 min	33,060	551
Quality control	3 weeks	7200	120
Total	–	71,025	1183.75

methodology for the selection, the annotation and the capture of the patients was refined during the 38 months of the collection of the cases, inconsistencies occurred, especially with the cases that were captured at the beginning of the project. The diagnosis of each case was retraced in the EHR with a methodology similar to the selection process. Cases with several co-morbidities were removed as the visual aspect can be altered strongly (e.g., in cardiac cardiac insufficiency). 21 cases were removed as their diagnoses were not reliably demonstrable or were mixed with co-morbidities. The values of several clinical parameters were corrected for 13 patients. When required, the entire set of clinical parameters was re-entered. Special care was taken for clinical parameters that are subject to change rapidly over time to use the parameters that are corresponding to the most acute manifestation of the disease and also as close as possible to the HRCT examination.

3.3. Total person-hour estimation

Estimations of the person-hours spent for the various tasks of the database construction are given in Table 6. The total person-hours spent are estimated at 1200. It is interesting to note that only 45.9% of that time was spent on really building the database (quality control included) whereas the remaining 54.1% were spent rejecting cases, capturing cases rejected by the quality control and re-entering clinical parameters. It is important to note that another

large amount of time spent for other tasks is not included in the 1200 person-hours, such as:

- selection of the histological diagnoses,
- selection of the clinical parameters to collect,
- iterative creation of the HTML forms,
- retrieval of the list of the potential ILD cases,
- creation of the database architecture,
- server installation and maintenance,
- development of the image annotation software,
- data conversions and anonymization,
- education of the personnel.

3.4. Data exploitation

The dataset served as reference for several studies on lung tissue analysis [38,39], case-based retrieval [6,18] and multi-modal medical data analysis [7]. Based on subsets of the database, results of computerized lung tissue categorization using techniques described in [18,40] are detailed in Tables 7 and 8. In addition, the performance of three-dimensional case-based retrieval based on histological diagnoses is given in Table 9 (see [6]). These performance results can be used as a baseline to compare with new methods on the same data set.

Table 7
Confusion matrix of the blockwise classification of lung tissue patterns in %. Global arithmetic and geometric means of 75.1% and 74.7% are obtained, respectively (*Source*: [18,40]). N_{vox} denotes the number of manually segmented voxels used for evaluation and N_{cases} the number of patients.

	Healthy	*Emphysema*	*Ground glass*	*Fibrosis*	*Micronodules*
Healthy	**78.1**	2.8	0.7	0.2	18.1
Emphysema	0.9	**70.1**	0	4.7	24.2
Ground glass	4.6	1.6	**76**	14.7	3.1
Fibrosis	2.3	1.9	17	**73.5**	5.3
Micronodules	13.7	1.8	2.2	6.7	**75.7**
N_{vox}	63,914	61,578	644,814	860,474	1,436,055
N_{cases}	7	5	21	28	10
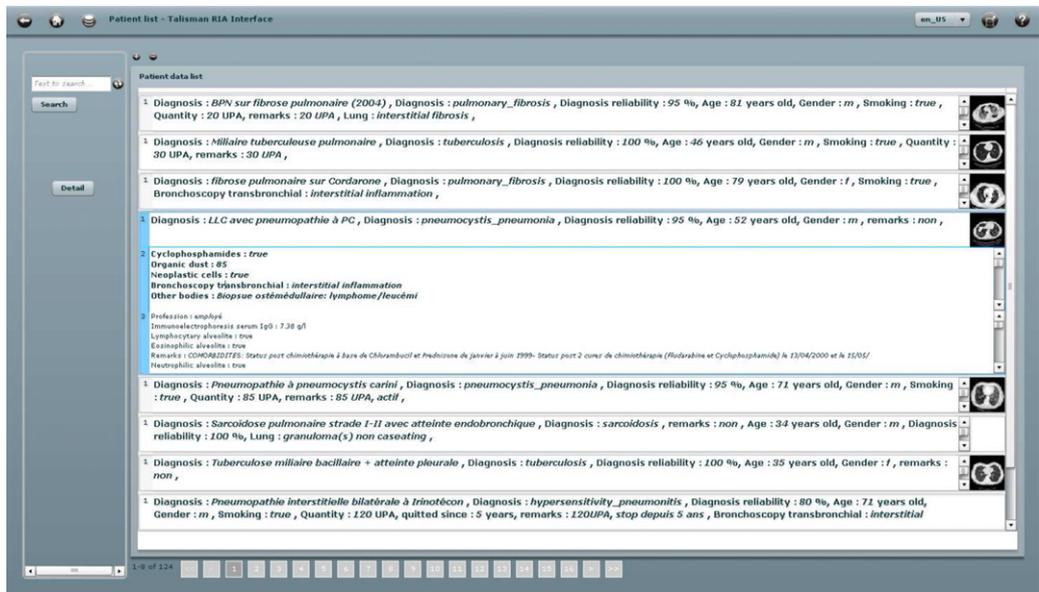
Fig. 10. List of cases returned by an exact text search. The user can preview the clinical parameters of the patients with three levels of importance. Levels of importance are varying based on the diagnosis.
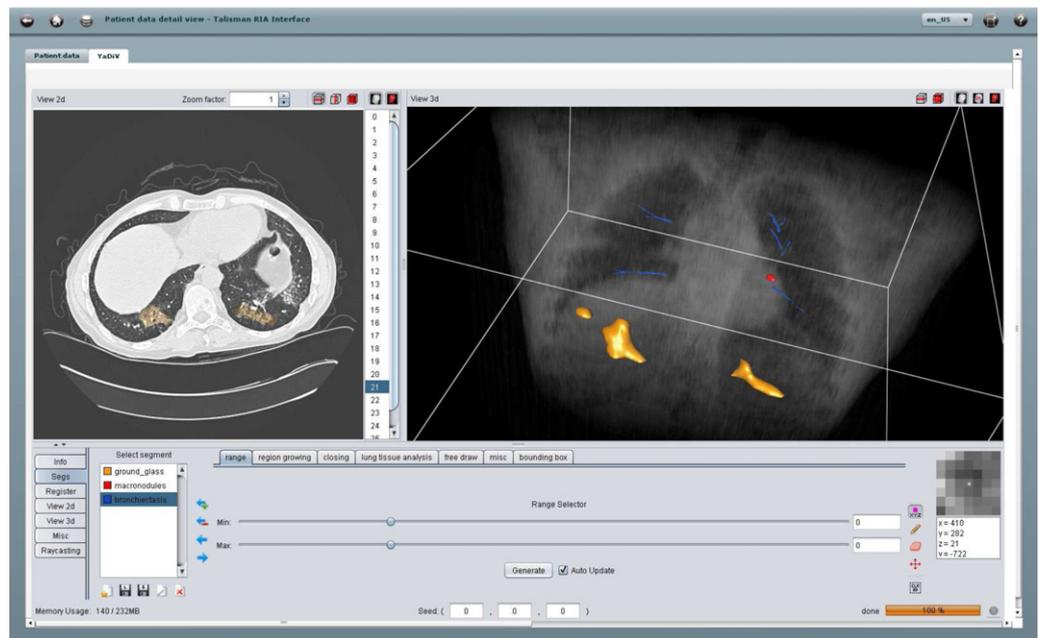
Fig. 11. Detailed view of a selected case: image series and annotations. 2D and 3D views are available to visualize the ROIs in YaDiV (*Source*: [41]).

Online consultation of the database for teaching is enabled by web-based interfaces shown in Figs. 10 and 11 [41]. The user can perform exact text search to browse the multimedia database.

Table 8
Performance measures of the blockwise classification of the lung tissue patterns (*Source*: [18,40]).

	Recall	Precision	F-measure	Accuracy
Healthy	78.4	78.2	78.3	91.3
Emphysema	89.6	70.2	78.7	92.4
Ground glass	79.2	76	77.6	91.2
Fibrosis	73.6	73.5	73.6	89.4
Micronodules	59.9	75.6	66.8	85

Table 9
Mean precisions at ranks 1, 5, 10 and at rank equal to the number of instances of the diagnosis N_r are computed based on the histological diagnosis of the retrieved cases (*Source*: [6]).

	P@1	P@5	P@10	P@N_r	N_r
PF	79.2	58.3	51.7	42.7	24
COP/BOOP	60	20	18	20	5
TB	71.4	48.6	34.3	42.9	7
PCP	25	20	10	25	4
HP	54.5	40	39.1	38	11
AIP	66.7	33.3	25.5	27.2	9
Sarcoidosis	100	66.6	52.2	56.8	9
Average/total	59.4	39.7	34.2	32.4	69
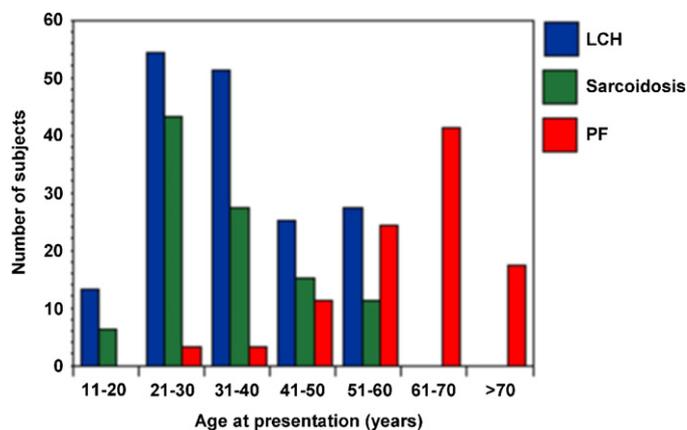
Fig. 12. Age at onset presentation in LCH, sarcoidosis and PF (*Source*: [20]).

4. Discussions

The distribution of the encountered diagnoses described in Table 4 and Fig. 7 is in accordance with the expected frequency in Europe [42]. A male predominance is observed as only 36.7% of the cases are female, which was also observed in [20,42,43]. The mean age of the patients with the respective diseases is shown in Table 4, and is in accordance with the expected values in the literature [20]. For instance, the distribution of the patient age with LCH, Sarcoidosis and PF shown in Fig. 12 is similar to the mean age of the corresponding diagnoses in Table 4. PF is found in patients over 50 [44] and the mean age of PF cases is 71.4 years in Table 4. The distribution of the diagnoses is very unbalanced where the most represented disease PF contains 40 patients and five diseases have only one case. The number of healthy control cases is also very low and is constituted by former ILD cases that have a histological proof of having recovered. These cases cannot truly be considered as control cases as the former ILD episode is related to abnormal values of some of the clinical parameters. As a consequence, distinguishing healthy and unhealthy tissue types is a hard task, that could be easier with really healthy patients.

The average time for completing the entire process of selecting, annotating the image series and capturing clinical data of a case was 3 h and 30 min, whereas a total 1183.75 person-hours was estimated for the contruction of the database. Thereby, an average of 9.25 person-hours per case captured was required. The 7 most represented types of lung tissue[12] in terms of number of ROIs allow describing a wide range of visual findings in HRCT image series associated with ILDs [23]. The distribution of annotated volumes of the various classes of lung tissue in Fig. 9 are not directly related to the distributions of the ROIs. This can be explained by observing the ratios of mean volume per ROI in Table 5. Based on the type of lung tissue, the size of the ROIs varies significantly. For instance, *micronodules* patterns found in cases of Miliary tuberculosis are diffusely distributed in the whole lung, which allowed to delineate very large regions. This is not the case for classes such as *macronodules* or *consolidation* for which the alterations of the lung parenchyma are narrowly localized. The annotation of the latter is comparatively more time-consuming compared to acquiring identical volumes of annotated tissue.

The annotations of the lung image regions are successively carried out by two radiologists with 15 and 20 years of experience. However, no assessment of the quality of the annotations by measuring inter-agreement measurement is carried out, which showed

to be an important source of errors in the interpretation process of HRCT image series showing diffuse parenchymal lung diseases in [13]. Nevertheless, the two radiologists have to agree on the delineated region and, thanks to the retrospective analysis, the annotations of the cases were based on the description of the image regions in the radiological report written by several radiologists from 2003 to 2008 at the HUG and a enabled cross-control of the interpretation of the image series.

Limitations occur in the representation of the variability of the class *healthy* as normal tissue was annotated in patients that have had an ILD in their medical history. Thereby, beyond continuing the collection of cases, future work is needed to add healthy cases in order to have a rich representation of healthy lung tissue and associated clinical parameters. This is challenging because healthy subjects that undergo an HRCT examination are very rare due to high cost and radiation dose delivery. On the other hand, the situation of difficult "healthy" tissue corresponds to the clinical reality particularly in emergency radiology.

5. Conclusion

In this paper, the construction steps and the content of a multimedia library of ILDs cases are described. The structured multimedia information served as basis for carrying out research on computerized image-based diagnosis aid for ILDs. So far, no similar datasets are publicly available for comparing pattern recognition techniques on ILDs. This situation makes it difficult to identify the best algorithms proposed in the literature to be further integrated into a CAD system used in a clinical environment. With the baseline system proposed a stable and reproducible standard technique allows to compare new techniques against a baseline in a controlled setting and thus reproducible results in a clear setting. The methodology proposed for the definition of the scope of the database (e.g., diseases and associated clinical parameters, imaging protocol), the selection of cases in an EHR as well as the annotation of medical image series constitutes a basis for collecting radiological data with the purpose of developing and evaluating computerized diagnostic aid.

References

[1] Horsch A, Prinz M, Schneider S, Sipilä O, Spinnler K, Vallée JP, et al. Establishing an international reference image database for research and development in medical image processing. Methods of Information in Medicine 2004;43(4):409–12.

[2] Hersh W, Müller H, Kalpathy-Cramer J, Kim E. Consolidating the ImageCLEF medical task test collection: 2005–2007. In: MUSCLE/ImageCLEF workshop. 2007. p. 31–9.

[3] Lowe HJ, Buchanan BG, Cooper GF, Vreis JK. Building a medical multimedia database system to integrate clinical information: an application of high-performance computing and communications technology. Bulletin of the Medical Library Association 1995;83(1):57–64.

[4] Doi K. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. Computerized Medical Imaging and Graphics 2007;31(4-5):198–211.

[5] Katsuragawa S, Doi K. Computer-aided diagnosis in chest radiography. Computerized Medical Imaging and Graphics 2007;31(4-5):212–23.

[6] Depeursinge A, Vargas A, Platon A, Geissbuhler A, Poletti PA, Müller H. 3D case-based retrieval for interstitial lung diseases. In: Proceedings of the MCBR-CDS 2009: medical content-based retrieval for clinical decision support. Lecture Notes in Computer Science (LNCS). Springer; 2010. p. 39–48.

[7] Depeursinge A, Racoceanu D, Iavindrasana J, Cohen G, Platon A, Poletti PA, et al. Fusing visual and clinical information for lung tissue classification

---

[12] i.e., *healthy*, *emphysema*, *ground glass*, *fibrosis*, *micronodules*, *consolidation* and *reticulation*.

in high-resolution computed tomography. Artificial Intelligence in Medicine 2010;50(1):13–21.

[8] Brodatz P. In: Textures: A photographic album for artists and designers. New York: Dover Publications; 1966.

[9] Friedman CP, Elstein AS, Wolf FM, Murphy GC, Franz TM, Heckerling PS, et al. Enhancement of clinician's diagnostic reasoning by computer-based consultation. Journal of the American Medical Association 1999;282(19):1851–6.

[10] Aisen AM, Broderick LS, Winer-Muram H, Brodley CE, Kak AC, Pavlopoulou C, et al. Automated storage and retrieval of thin-section CT images to assist diagnosis: system description and preliminary assessment. Radiology 2003;228(1):265–70.

[11] Müller H, Michoux N, Bandon D, Geissbuhler A. A review of content-based image retrieval systems in medicine—clinical benefits and future directions. International Journal of Medical Informatics 2004;73(1):1–23.

[12] Müller H, Rosset A, Vallée JP, Terrier F, Geissbuhler A. A reference data set for the evaluation of medical image retrieval systems. Computerized Medical Imaging and Graphics 2004;28(6):295–305.

[13] Aziz ZA, Wells AU, Hansell DM, Bain GA, Copley SJ, Desai SR, et al. HRCT diagnosis of diffuse parenchymal lung disease: inter-observer variation. Thorax 2004;59(6):506–11.

[14] Armato SG, McNitt-Gray MF, Reeves AP, Meyer CR, McLennan G, Aberle DR, et al. The lung image database consortium (LIDC): an evaluation of radiologist variability in the identification of lung nodules on CT scans. Academic Radiology 2007;14(11):1409–21.

[15] Rajasekaran H, Lo Iacono L, Hasselmeyer P, Fingberg J, Summers P, Benkner S, et al. Aneurist—towards a system architecture for advanced disease management through integration of heterogeneous data, computing, and complex processing services. In: Proceedings of the 21th IEEE symposium on computer-based medical systems (CBMS). Finland: Jyväskylä; 2008. p. 361–6.

[16] Roduit N, Meyer R, Bijlenga P, Geissbuhler A. Image depersonalization: a new efficient approach. In: Proceedings of the RSNA 2009: 95th scientific assembly and annual meeting of the Radiological Society of North America. RSNA; 2009.

[17] Elger B, Iavindrasana J, Lo Iacono L, Müller H, Roduit N, Summers P, et al. Strategies for health data exchange for secondary, cross-institutional clinical research. Computer Methods and Programs in Biomedicine 2010;99(3):230–51.

[18] Depeursinge A. Affine-invariant texture analysis and retrieval of 3D medical images with clinical context integration. PhD thesis. University of Geneva; 2010.

[19] Depeursinge A, Iavindrasana J, Andriambololoniaina H, Poletti PA, Platon A, Geissbuhler A, et al. A framework for diagnosing interstitial lung diseases in HRCT: the TALISMAN project. Swiss Medical Informatics 2008;64:17–20.

[20] King TE. In: Basow Denise S, editor. Approach to the adult with interstitial lung disease. Waltham, MA: UpToDate; 2008.

[21] Flaherty KR, King TE, Raghu G, Lynch JP, Colby TV, Travis WD, et al. Idiopathic interstitial pneumonia: what is the effect of a multidisciplinary approach to diagnosis? American Journal of Respiratory and Critical Care Medicine 2004;170(8):904–10.

[22] Stark P. In: Basow Denise S, editor. High resolution computed tomography of the lungs. Waltham, MA: UpToDate; 2008.

[23] Webb WR, Müller NL, Naidich DP. In: High-resolution CT of the lung. Philadelphia, USA: Lippincott Williams & Wilkins; 2001.

[24] Müller NL. Clinical value of high-resolution CT in chronic diffuse lung disease. American Journal of Roentgenology 1991;157(6):1163–70.

[25] Lutterbey G, Grohé C, Gieseke J, von Falkenhausen M, Morakkabati N, Wattjes MP, et al. Initial experience with lung-MRI at 3.0 T: comparison with CT and clinical data in the evaluation of interstitial lung disease activity. European journal of Radiology 2007;61(2):256–61.

[26] Souza CA, Müller NL, Flint J, Wright JL, Churg A. Idiopathic pulmonary fibrosis: spectrum of high-resolution CT findings. American Journal of Roentgenology 2005;185(6):1531–9.

[27] Doi K. Current status and future potential of computer-aided diagnosis in medical imaging. British Journal of Radiology 2005;78:3–19.

[28] Kakinuma R, Ohmatsu H, Kaneko M, Eguchi K, Naruke T, Nagai K, et al. Detection failures in spiral CT screening for lung cancer: analysis of CT findings. Radiology 1999;212(1):61–6.

[29] Nowers K, Rasband JD, Berges G, Gosselin M. Approach to ground-glass opacification of the lung. Seminars in Ultrasound, CT, and MRI 2002;23(4):302–23. Diffuse Lung Disease.

[30] Armato SG, McLennan G, McNitt-Gray MF, Meyer CR, Yankelevitz D, Aberle DR, et al. Lung image database consortium: developing a resource for the medical imaging research community. Radiology 2004;232(3):739–48.

[31] McNitt-Gray MF, Armato SG, Meyer CR, Reeves AP, McLennan G, Pais RC, et al. The lung image database consortium (LIDC) data collection process for nodule detection and annotation. Academic Radiology 2007;14(12):1464–74.

[32] Reeves AP, Biancardi AM, Yankelevitz DF, Fotin S, Keller BM, Jirapatnakul AC, et al. A public image database to support research in computer aided diagnosis. In: Proceedings of the 31st annual international conference of the IEEE Engineering in Medicine and Biology Society. 2009. p. 3715–8.

[33] Holmes III DR, Bartholmai BJ, Karwoski RA, Zavaletta V, Robb RA. The lung tissue research consortium: an extensive open database containing histological, clinical, and radiological data to study chronic lung disease. Insight Journal 2006.

[34] Karwoski RA, Bartholmai B, Zavaletta VA, Holmes D, Robb RA. Processing of CT images for analysis of diffuse lung disease in the lung tissue research consortium. In: Hu XP, Clough AV, editors. Medical imaging 2008: physiology, function, and structure from medical images, vol. 6916. SPIE; 2008, 691614–691614-9.

[35] Rudrapatna M, Sowmya A, Zrimec T, Wilson P, Kossoff G, Lucas P, et al. LMIK—learning medical image knowledge: an internet-based medical image knowledge acquisition framework. In: Santini S, Schettini R, editors. Internet imaging V, vol. 5304. SPIE; 2003. p. 307–18.

[36] Hidki A, Müller H, Depeursinge A, Poletti PA, Geissbuhler A. Putting the image into perspective: the need for domain knowledge when performing image-based diagnostic aid. In: Proceedings of the Swiss conference on medical informatics (SSIM 2006). 2006.

[37] Müller H, Marquis S, Cohen G, Geissbuhler A. Lung CT analysis and retrieval as a diagnostic aid. In: Engelbrecht R, Geissbuhler A, Lovis C, Mihalas G, editors. Proceedings of the medical informatics Europe conference (MIE 2005), vol. 116. 2005. p. 453–8.

[38] Depeursinge A, Sage D, Hidki A, Platon A, Poletti PA, Unser M, et al. Lung tissue classification using wavelet frames. In: Proceedings of the 29th annual international conference of the IEEE Engineering in Medicine and Biology Society, 2007 (EMBS 2007). Lyon, France: IEEE Computer Society; 2007. p. 6259–62.

[39] Depeursinge A, Van De Ville D, Unser M, Müller H. Lung tissue analysis using isotropic polyharmonic B-spline wavelets. In: Proceedings of the MICCAI 2008 workshop on pulmonary image analysis. 2008. p. 125–34.

[40] Depeursinge A, Zrimec T, Busayarat S, Müller H. 3D lung image retrieval using localized features. In: Proceedings of the medical imaging 2011: computer-aided diagnosis, vol. 7963. SPIE; 2011, 79632E-1.

[41] Depeursinge A, Platon A, Geissbuhler A, Poletti PA, Müller H. Case-based lung image categorization and retrieval for interstitial lung diseases: clinical workflows. International Journal of Computer Assisted Radiology and Surgery; in press. Available online at: http://www.springerlink.com/content/4146pm446g311455/.

[42] Thomeer MJ, Costabel U, Rizzato G, Poletti V, Demedts M. Comparison of registries of interstitial lung diseases in three European countries. European Respiratory Journal 2001;18(32):114–8.

[43] Chiyo M, Sekine Y, Iwata T, Tatsumi K, Yasufuku K, Iyoda A, et al. Impact of interstitial lung disease on surgical morbidity and mortality for lung cancer: analyses of short-term and long-term outcomes. Journal of Thoracic and Cardiovascular Surgery 2003;126(4):1141–6.

[44] Wade JF, King TE. Infiltrative and interstitial lung disease in the elderly patient. Clinics in Chest Medicine 1993;14(3):501–21.