Readme for VQA-RAD, a dataset of visual questions and answers in radiology

1. GENERAL INFORMATION

Title of dataset: VQA-RAD

Dina Demner-Fushman (ddemner@mail.nih.gov)

Lister Hill National Center for Biomedical Communications, National Library of Medicine, Bethesda, MD, USA

2.0 DATA AND FILE OVERVIEW

Link: Lau, Jason. Open Science Framework. doi: <https://osf.io/89kps/?view_only=521f76b347b146ccbe85ee24396849c8> (2018)

|  |  |
| --- | --- |
| VQA\_RAD Dataset.json | VQA-RAD full dataset of question and answers referencing images. File formats include XML, JSON, and EXCEL. Additional metadata includes categories and labels, see DATA SPECIFIC INFORMATION |
| VQA\_RAD Dataset.xml |
| VQA\_RAD Dataset.xlsx |
| VQA\_RAD Image folder | Folder of 315 radiological images referenced from question and answers. Images are varying sizes, all are .jpeg. |
|  |  |

3.0 DATA SPECIFIC INFORMATION

VQA\_RAD Dataset 2018\_06\_011 for JSON, XML, and Excel formats

|  |  |
| --- | --- |
| Number of variables | 14 |
| Numbers of rows | 2248 |

|  |  |  |
| --- | --- | --- |
| Variable | Description | Section |
| Image\_name | Name of image to “VQA\_RAD Images” file | 4.1 |
| Image\_case\_url | Image link to MedPix® case which includes original image, caption, and other contextual information | 4.1 |
| Image\_organ | Type of image organ system e.g. Head, Chest, Abdomen | 4.1 |
|  |  |  |
| question | Visual question about image | 4.2 |
| Qid | Unique identifier for all free-form and paraphrased questions |  |
| Phrase\_type | Whether question is original free-form question or rephrased from another question  Freeform = original question  Para = rephrased from another question  Test\_freeform = original question used for test data  Test\_paraphrase = rephrased questions of the test\_freeform | 4.2 |
| Question\_type | Type of question:  MODALITY  PLANE  ORGAN (Organ System)  ABN (Abnormality)  PRES (Object/Condition Presence)  POS (Positional Reasoning)  COLOR  SIZE  ATTRIB (Attribute Other)  COUNT (Counting)  Other | 4.3 |
|  |  |  |
| Answer | Answer to the question | 4.2 |
| Answer\_type | type of answer, e.g. closed-ended, open-ended | 4.4 |
|  |  |  |
| Evaluation | Whether question-answer pair was clinically evaluated by a 2nd clinician, e.g evaluated = two clinical annotators reviewed image and QA pair, not evaluated = one clinical annotator | 4.5 |
| Question\_relation | Relationship between linked question-answer pairs e.g.  Strict agreement  Loose agreement   * Inversion * Conversion * Subsumption   Not similar | 4.5 |
| Qid\_linked\_id | Unique identifier for every pair of free-form and paraphrased questions that can be used to link original and rephrasing |  |
| Question\_rephrase | Rephrasing of ‘question’, can be freeform or para, linked through qid\_linked\_id | 4.2 |
| Question\_frame | Rephrasing of ‘question’ following a templated structure | 4.2 |

VQA\_RAD Image folder

|  |  |
| --- | --- |
| Number of images | 315 |
| Format | .jpeg |

4.0 METHODOLOGICAL INFORMATION

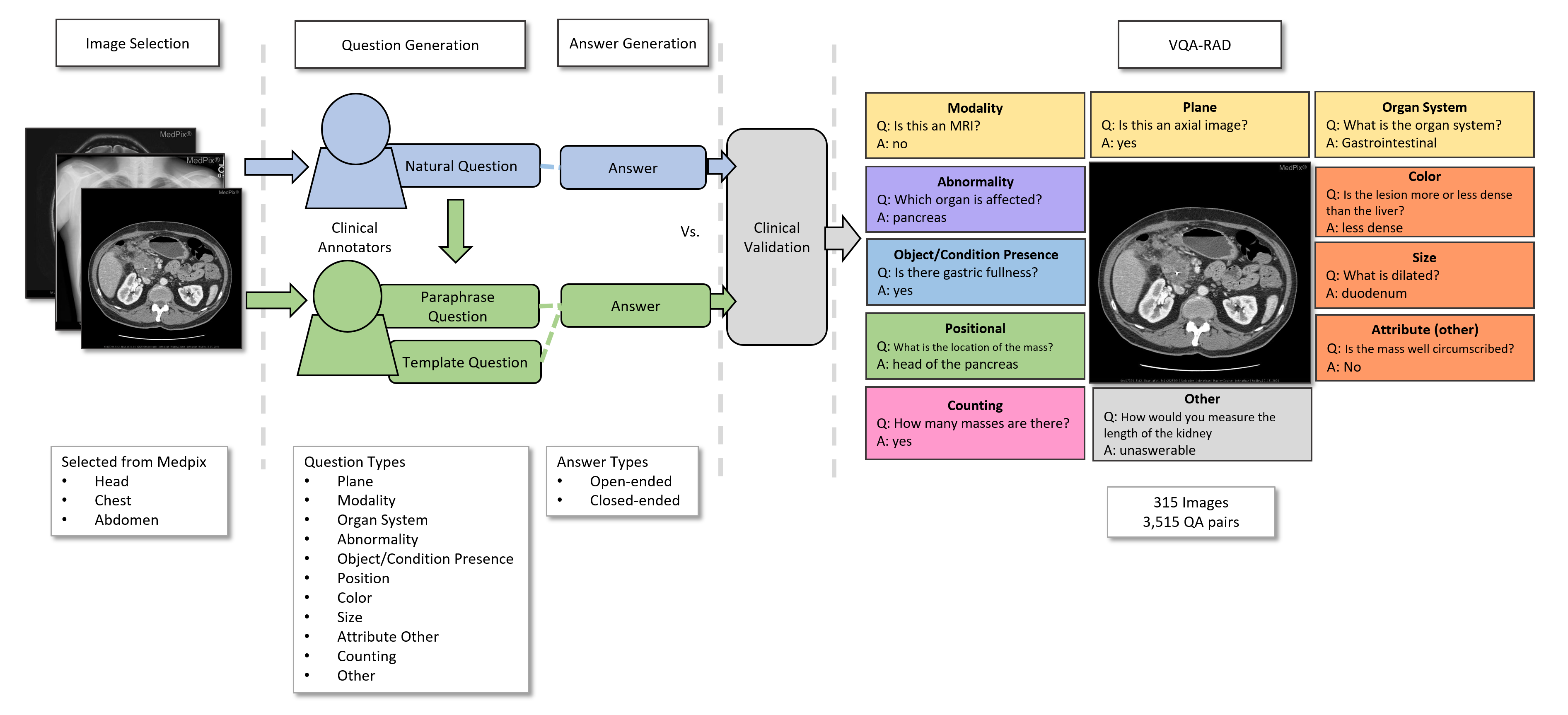


Figure 1. Flow Diagram of VQA-RAD build

4.1 IMAGE SELECTION

We sampled images from teaching cases in MedPix, https://medpix.nlm.nih.gov/, an open-access database of radiology images and teaching cases. Our sampling criteria were as follows: (1) Only one image for each teaching case, so that all images represented unique patients. (2) All images are sharp enough to identify individual structures. (3) Images are clean of radiology markings, such as arrows or circles. (4) Images have captions that correspond to the image and are detailed enough to describe at least one structure. Captions include plane, modality, and image findings that were generated and reviewed by expert radiologists. In total, we selected 104 head axial CTs or MRIs, 107 chest x-rays, and 104 abdominal axial CTs. The balanced distribution from head, chest, and abdomen should help determine if visual questions differ for each organ system and if the algorithms perform differently on different regions.

4.2 QUESTION AND ANSWER GENERATION

Questions and answers were generated by 15 volunteer clinical trainees using a web-interface developed for collecting the questions and the answers. All participants had completed the core rotations of medical school, which typically occurs during the 3rd year of school and exposes students to major fields of medicine such as surgery, internal medicine, neurology, etc. This ensures that all participants have basic clinical radiology reading skills and were exposed to a variety of settings where radiology was vital to the management of patients. Our participants had training from different regions of the U.S. and have interests in different specialties including radiology, orthopedics, family medicine, etc.

Participants generated questions and answers in a two-part evaluation (shown in Figure 1) from December 2017 to April 2018. Each participant reviewed at least 40 randomized images. For the first 20 images, participants provided “**free-form**” questions and answers without any restrictions. We instructed participants to create “free-form” question about the images by phrasing them in a natural way as if they are asking a colleague or another physician. The image alone had to be sufficient to answer the question and there should only be a single correct answer. We asked that answers to the visual questions be based off their level of knowledge. Since many of the participants were still in medical training, we provided captions with some image findings, plane, and modality information to provide additional ground truth reassurance.

For next 20 images, participants were randomly paired and given another participant’s images and questions. They were asked to generate “**paraphrased**” and “**framed**” questions based off the given “free-form” questions with corresponding image and caption. We asked the participants to paraphrase the question in a natural way and generate an answer that agreed with both the original and the paraphrased questions.

Participants generated “framed” questions by finding the closest question structure from a list of templates and filling in the blank spaces to retain the answer to the original questions.

4.3 QUESTION TYPES

|  |  |
| --- | --- |
| Question Type | Description |
| **Modality** | How an image is taken – CT, x-ray, T2 weighted MRI, etc. |
| **Plane** | Orientation of an image slicing through the body – axial, sagittal, coronal |
| **Organ System** | Categorization that connects anatomical structures with pathophysiology, diagnosis, and treatment – pulmonary, cardiac, musculoskeletal system |
| **Abnormality** | Normalcy of an image or object. For example, “is there something wrong with the image?” or “What is abnormal about the lung?”, “Does the liver look  normal?” |
| **Object/Condition Presence** | Objects could be normal structures like organs or body parts but could also be abnormal objects such as masses or lesions. Clinicians may refer to the presence of conditions in an image or patient – fractures, midline shift, infarction |
| **Positional reasoning** | position or location of an object or organ, including what side of a patient, in respect to the image borders, or relative to other objects in the image |
| **Color** | signal intensity including enhancement or opaqueness |
| **Size** | measurement of size of an object, e.g., enlargement, atrophy |
| **Attribute Other** | other types of description questions |
| **Counting** | focusing on a quantity of objects, e.g., number of lesions |
| **Other** | catch-all categorization for questions that do not fall into the previous categories |

We identify three categories, modality, plane, and organ system questions, that contribute to baseline knowledge for every radiological image. **Modality** questions, which refer to how an image is taken (CT with contrast, x-ray, T2 weighted MRI, or etc), help give context for identifying white and black structures. Active bleeding can appear as a white mass on a CT scan while only a dull grey on MRI. **Plane** questions, which refer to the orientation of the image slicing through the body, helps to understand anatomical structures. **Organ system** is a subjective categorization that depends on the clinical context. However, it is an important concept frequently taught to all clinicians to connect pieces of anatomy with pathophysiology, diagnoses, and treatment. An image of a chest can contain multiple organ systems such as pulmonary, cardiac, gastrointestinal, or orthopedic systems. These boundaries help decision making, for example, a mass near the heart could represent a pneumonia that does not directly affect the heart itself despite the proximity. We consider these three categorizations as baseline question types because most clinicians, regardless of experience, have some understanding while the public knowledge may not.

**Abnormality** questions ask about the normalcy of an image or object. For example, “is there something wrong with the image?” or “What is abnormal about the lung?”, “Does the liver look

normal?”

**Presence** questions represent questions focusing on the presence of an object or condition. Objects could be normal structures like organs or body parts but could also be abnormal objects such as masses or lesions. Clinicians may refer to the presence of conditions in an image or patient; however, they are more difficult to delineate. For example, ‘fracture’ is condition that can happen to a bone which one might say “a fractured femur”, “a fracture in the femur”, or “the femur has a fracture”. All of these linguistic variations are still referring to the presence of a fracture in the image. Other examples of common conditions include ‘midline shift’, ‘pneumothorax’, and ‘infarction’.

**Positional reasoning** questions focus on the position or location of an object or organ, including what side of a patient. Easily labeled questions are “Where is the mass?” or “Is the lesion located on the left?”. Some difficulty in typing arose between positional and presence questions. The question, “Is the mass in the left lung?” is both asking the presence of a mass and the position of the mass on the left. In this dataset, we chose to label such questions as presence questions rather than positional though ideally these questions would have both labels.

Attribute questions include color, size, and other attribute which are questions that focus on the description of an object rather than its position or presence. Since most radiological images are grey, **Color** questions refer to the signal intensity including enhancement or opaqueness. **Size** questions are ones that need a measurement of size of an object to answer it. These included words like enlargement, atrophy, dilation. These categories are important to separate because unique tools may be needed to answer these questions, such as size questions needing a reference of normal size or color questions needing to normalize the signal intensities. Since there may be other common attributes size and color, the **Attribute Other** question type is tagged for other types of description questions.

**Counting** questions are fairly straight forward to label. Any question focusing on a quantity of objects. Care is taken to ensure that answers are only based on a single image. Some radiological captions may describe a series of images showing multiple slices of the body, but we limit this dataset to what can be answered with one image.

**Other** question type is a catch-all categorization for questions that do not fall into the previous categories. Examples include questions that require epidemiological knowledge or next step treatments.

4.4 ANSWER TYPES

|  |  |
| --- | --- |
| Answer Type | Description |
| **Close-ended** | yes/no and other limited choices. For example, “Is the mass on the left or right?” |
| **Open-ended** | Do not have a limited question structure and could have multiple correct answers |

Answer types are labeled after the evaluation completion. **Closed-ended** and **open-ended** answers are the only categories we used. Closed-ended answer include yes/no and other limited choices. For example, “Is the mass on the left or right?” is a closed-ended structure. Open-ended answers did not have a limited question structure and could have multiple correct answers.

4.5 QUESTION ANSWER VALIDATION

After completion of the evaluations, we used several methods to validate questions answer pairs and question types. During the paraphrasing part of the evaluation, participants answered another person’s questions. The answers could have strict or loose agreement. We defined **strict agreement** when the question and answer format and topic were the same. In loose agreement, the topic of the questions is the same or similar even though the answers may differ. Three subcategories of loose agreement are defined: inversion, conversion, and subsuming.

Examples of each as follows:

**Inversion:** Q1:”Are there abnormal findings in the lower lung fields?” is a negation of Q2:” “Are the lower lung fields normal?”

**Conversion**: Q1:“How would you describe the abnormalities?” is open-ended while Q2:“Are the lesions ring-enhancing?” is closed-ended

**Subsumption:** Q1: “Is the heart seen in the image?” subsumes Q2:“is the heart seen on the left?”

Questions are considered ‘**evaluated**’ when they were reviewed by two annotators. Disagreements in answers are resolved by research team consensus and expert radiologist review. Questions are labeled as ‘**not evaluated**’ if they are not reviewed by a second participant or the paraphrased question is not similar enough to be used as validation. Both the evaluated and not evaluated questions are used as part of the test and training set.

We validated question types assigned by the participants. Final categorization was determined through consensus with the research team to resolve disagreements.

5.0 TRAINING AND TEST SET

To demonstrate a use case of VQA-RAD, we created a training and test set by randomly sampling the free-form questions and then matching the corresponding paraphrased questions. The resultant test set is composed of 300 randomly chosen free-form questions and 151 corresponding paraphrased questions. We used the remainder of the free-form and paraphrased questions as the training set.

Other training and test sets can be created using the free-form, paraphrasing, and framed questions. Since these question sets can share a single answer, we recommend isolating a phrase type (phrase\_type), randomly selecting the proportion of questions, and finding matched questions using the qid\_linked\_id and the question\_frame variables. This method can limit bias that may occur if, for example, a free-form question is used in the training and the paired paraphrased question is in the test set.

6.0 DATASET CHARACTERISTICS

Figure 2. Closed vs Open Ended Questions and Breakdown of different types (free form questions only) Certain questions types more likely to be open-ended: positional, counting questions and other.

Figure 3. Question type per Image Organ type (free form questions only). Most HEAD questions about color/signal intensity. Most CHEST questions about size. Fewer positional questions about the ABDOMEN than other image organs.

DISTINCT WORD DISTRIBUTION AND FREQUENCY (free-form questions only)

Tables for each question type and answer type showing total number of questions, median number of words per question, total number of words, and number of distinct words. Distinct words determined by tokenizing sentences and making all words lowercase. Also shown are top 10 most common words with percent frequency word appears for category. Stop words removed and bold words demonstrate words only appearing in the 10 top of question type (i.e. MRI, weighted, and IV only appear in Modality Questions).

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | MODALITY | |  |  | PLANE |  |  |  | ORGAN |  |  |  |
|  | closed |  | open |  | closed |  | open |  | closed |  | open |  |
| #questions | 79 |  | 67 |  | 55 |  | 47 |  | 15 |  | 36 |  |
| median question length (#words) | 6 |  | 6 |  | 6 |  | 6 |  | 7 |  | 7.5 |  |
| #words total | 487 |  | 439 |  | 342 |  | 296 |  | 103 |  | 265 |  |
| #distinct words | 94 |  | 44 |  | 55 |  | 40 |  | 38 |  | 66 |  |
| 1 | CONTRAST | 48.1% | IMAGE | 68.7% | IMAGE | 40.0% | **PLANE** | 76.6% | IMAGE | 60.0% | ORGAN | 66.7% |
| 2 | IMAGE | 36.7% | **MODALITY** | 52.2% | **PLANE** | 40.0% | IMAGE | 63.8% | **SYSTEM** | 46.7% | **SYSTEM** | 50.0% |
| 3 | **CT** | 31.6% | TYPE | 50.7% | **AXIAL** | 36.4% | TAKEN | 46.8% | BRAIN | 33.3% | IMAGE | 47.2% |
| 4 | **MRI** | 19.0% | IMAGING | 28.4% | **PA** | 23.6% | WHICH | 34.0% | CHEST | 20.0% | IMAGED | 33.3% |
| 5 | **WEIGHTED** | 11.4% | TAKEN | 16.4% | **FILM** | 16.4% | ABOVE | 12.8% | DISPLAY | 13.3% | PART | 11.1% |
| 6 | **IV** | 11.4% | **MRI** | 13.4% | TAKEN | 12.7% | ACQUIRED | 6.4% | PATHOLOGY | 13.3% | ABOVE | 8.3% |
| 7 | GIVEN | 10.1% | CONTRAST | 7.5% | **AP** | 12.7% | BODY | 4.3% | PULMONARY | 13.3% | BODY | 8.3% |
| 8 | PATIENT | 7.6% | KIND | 7.5% | **CORONAL** | 10.9% | **CUT** | 4.3% | STUDY | 13.3% | EVALUATED | 8.3% |
| 9 | **SCAN** | 7.6% | ABOVE | 7.5% | BRAIN | 9.1% | **FILM** | 4.3% | ABDOMEN | 6.7% | PRIMARILY | 8.3% |
| 10 | TAKEN | 7.6% | ACQUIRE | 6.0% | **SAGGITAL** | 7.3% | WHERE | 4.3% | **CARDIOVASCULAR** | 6.7% | SHOWN | 8.3% |

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | ABNORMALITY | |  |  | PRESENCE | |  |  | POSITION |  |  |  |
|  | closed |  | open |  | closed |  | open |  | closed |  | open |  |
| #questions | 78 |  | 32 |  | 379 |  | 104 |  | 19 |  | 154 |  |
| median question length (#words) | 5 |  | 6 |  | 6 |  | 8 |  | 9 |  | 7 |  |
| #words total | 441 |  | 204 |  | 2428 |  | 899 |  | 185 |  | 1032 |  |
| #distinct words | 112 |  | 69 |  | 395 |  | 246 |  | 69 |  | 202 |  |
| 1 | **NORMAL** | 52.6% | **PATHOLOGY** | 34.4% | PRESENT | 15.0% | IMAGE | 26.9% | LEFT | 31.6% | **WHERE** | 32.5% |
| 2 | IMAGE | 29.5% | IMAGE | 28.1% | IMAGE | 10.6% | LEFT | 10.6% | **LOCATED** | 31.6% | **WHICH** | 27.9% |
| 3 | ABNORMAL | 23.1% | ABNORMAL | 9.4% | **EVIDENCE** | 9.5% | RIGHT | 9.6% | LUNG | 26.3% | **LOCATED** | 20.8% |
| 4 | LIVER | 20.5% | **ABNORMALITY** | 9.4% | AIR | 6.3% | ORGAN | 7.7% | OPACITIES | 21.1% | LESION | 16.9% |
| 5 | ABNORMALITIES | 9.0% | **INVOLVED** | 9.4% | MASS | 5.8% | MASS | 5.8% | RIGHT | 21.1% | MASS | 13.6% |
| 6 | FINDINGS | 7.7% | LUNG | 9.4% | **FRACTURE** | 5.3% | SIDE | 5.8% | SIDE | 21.1% | SIDE | 11.7% |
| 7 | AIR | 6.4% | ABNORMALITIES | 6.3% | LEFT | 5.3% | ANTERIOR | 4.8% | CONTRAST | 15.8% | IMAGE | 9.7% |
| 8 | LUNGS | 6.4% | HAPPENING | 6.3% | **PNEUMOTHORAX** | 4.7% | BRAIN | 4.8% | LESION | 15.8% | ABNORMALITY | 9.1% |
| 9 | **ABNORMALITY** | 5.1% | LESION | 6.3% | RIGHT | 4.7% | BRIGHT | 4.8% | AORTA | 10.5% | BRAIN | 9.1% |
| 10 | BRAIN | 5.1% | PANCREAS | 6.3% | BOWEL | 4.2% | **HYPERDENSITIES** | 4.8% | BOWELS | 10.5% | LOBE | 4.5% |

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | COLOR |  |  |  | SIZE |  |  |  | ATTRIBUTE (OTHER) | |  |  |
|  | closed |  | open |  | closed |  | open |  | closed |  | open |  |
| #questions | 25 |  | 7 |  | 91 |  | 10 |  | 29 |  | 17 |  |
| median question length (#words) | 6 |  | 7 |  | 5 |  | 7 |  | 6 |  | 6 |  |
| #words total | 171 |  | 56 |  | 502 |  | 67 |  | 177 |  | 108 |  |
| #distinct words | 69 |  | 26 |  | 48 |  | 22 |  | 71 |  | 41 |  |
| 1 | LESION | 24.0% | **INTENSITY** | 42.9% | **ENLARGED** | 28.6% | MASS | 70.0% | MASS | 34.5% | DESCRIBE | 58.8% |
| 2 | MASS | 20.0% | ABNORMALITY | 28.6% | HEART | 25.3% | LESION | 40.0% | LESION | 20.7% | LESION | 29.4% |
| 3 | ENHANCING | 16.0% | DENSITY | 28.6% | NORMAL | 11.0% | **SIZE** | 40.0% | **CYSTIC** | 13.8% | ABNORMAL | 11.8% |
| 4 | **HYPER** | 16.0% | DESCRIBE | 28.6% | **SIZE** | 9.9% | **LARGE** | 30.0% | ENHANCING | 10.3% | IMAGE | 11.8% |
| 5 | MORE | 16.0% | LESION | 28.6% | **DILATED** | 8.8% | **BIG** | 20.0% | **HOMOGENEOUS** | 10.3% | MASS | 11.8% |
| 6 | THAN | 16.0% | **SIGNAL** | 28.6% | CARDIAC | 7.7% | **CM** | 10.0% | **RING** | 10.3% | ABNORMALITIES | 5.9% |
| 7 | ABNORMALITY | 12.0% | **AREA** | 14.3% | AORTA | 6.6% | DENSITY | 10.0% | **CIRCUMSCRIBED** | 6.9% | ADJECTIVE | 5.9% |
| 8 | ATTENUATED | 12.0% | **BLACK** | 14.3% | **CARDIOMEGALY** | 6.6% | DESCRIBE | 10.0% | **CONTOUR** | 6.9% | APPENDIX | 5.9% |
| 9 | CONTRAST | 12.0% | **CENTRAL** | 14.3% | **ENLARGEMENT** | 6.6% | LOCATED | 10.0% | **FLATTENED** | 6.9% | ARTERY | 5.9% |
| 10 | DENSE | 12.0% | **COLOR** | 14.3% | SILHOUETTE | 6.6% | **QUADRANT** | 10.0% | **HEMIDIAPHRAGMS** | 6.9% | **BORDERS** | 5.9% |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | COUNT |  |  |  | OTHER |  |  |  |
|  | closed |  | open |  | closed |  | open |  |
| #questions | 6 |  | 9 |  | 34 |  | 51 |  |
| median question length (#words) | 10.5 |  | 7 |  | 8 |  | 9 |  |
| #words total | 57 |  | 63 |  | 274 |  | 485 |  |
| #distinct words | 41 |  | 23 |  | 130 |  | 205 |  |
| 1 | JUST | 33.3% | **MANY** | 100.0% | PATIENT | 29.4% | IMAGE | 19.6% |
| 2 | MORE | 33.3% | IMAGE | 55.6% | IMAGE | 23.5% | PATHOLOGY | 11.8% |
| 3 | **MULTIPLE** | 33.3% | MASSES | 33.3% | INJURY | 11.8% | PATIENT | 11.8% |
| 4 | **ONE** | 33.3% | KIDNEYS | 22.2% | DIAGNOSIS | 8.8% | LEFT | 9.8% |
| 5 | THAN | 33.3% | LESIONS | 22.2% | HEART | 8.8% | **SUGGEST** | 9.8% |
| 6 | **1** | 16.7% | ENHANCING | 11.1% | LYING | 8.8% | **WHY** | 9.8% |
| 7 | **2** | 16.7% | FOUND | 11.1% | MASS | 8.8% | LIKELY | 7.8% |
| 8 | **5** | 16.7% | GALLSTONES | 11.1% | PROCESS | 8.8% | CXR | 5.9% |
| 9 | **8** | 16.7% | **IDENTIFIED** | 11.1% | SUPINE | 8.8% | MASS | 5.9% |
| 10 | **>1** | 16.7% | **INSTANCES** | 11.1% | SUSPECT | 8.8% | MOST | 5.9% |