2012 Cairo International Biomedical Engineering Conference (CIBEC)
Cairo, Egypt, December 20-21, 2012

On the Comparability of Medical Image Measurements in Cancer Clinical Trials

Omar El Gazzar, Michael Onken, Marco Eichelberg, Andreas Hein
OFFIS – Institute for Information Technology
Escherweg 2, 26121 Oldenburg, Germany
omar.el.gazzar@offis.de

***Abstract*—Biomarkers detectable by imaging methods are increasingly used as surrogate endpoints to establish the severity of a disease to evaluate the efficacy of a treatment method or chemotherapy drug in cancer clinical trials. In therapeutic cancer clinical trials, the tumors are measured from medical images such as CT or MR at a baseline time point and the treatment response assessment is followed-up on a frequent basis at subsequent time intervals. In this paper, we will try to evaluate to which degree the measurements from medical images are comparable for clinical trials by specifying sources of uncertainties of image measurements in order to support both clinical trial investigators by giving them confidence in their interpretation of the measurements results and implementers to follow the best-practice guidelines to be taken into consideration to minimize sources of such variability.**

***Index Terms*—clinical trials, quality assurance, DICOM, medical images, RECIST.**

I. Introduction

A clinical trial or clinical study is defined as "a research investigation in human subjects that is designed to assess the safety and efficacy of a biomedical intervention (drug, treatment, or device) or new ways of using a known drug, treatment, or device" [1]. If the clinical trial is randomized, the selected subjects are randomly assigned to different treatment arms in the study in order to fairly compare between the evaluated treatment methods. The primary clinical endpoint in clinical trials is the patient survival rate. However, imaging biomarkers are increasingly used as surrogate endpoints for evaluating the treatment response. A surrogate endpoint is "an endpoint obtained more quickly, at lower cost, or less invasively than the true clinical endpoint of interest" [2]. For example, a surrogate endpoint in lung cancer clinical trial that is evaluated using imaging techniques is the tumor size growth or shrinkage. Tumor size can be quantified using 2-D image measurements or 3-D volumetric measurements obtained on a frequent basis to evaluate the treatment.

There are some constraints facing clinical trial investigators while using imaging techniques for quantitative assessment of the evaluated tumor treatment response. One of the main constraints in using imaging methods is the uncertainty of the obtained image measurements and their comparability between different subjects (inter-variability) and for the same subject at different time points (intra-variability).

II. Objectives

Our objective is to outline the factors affecting the comparability of acquired medical images and their usability for performing consistent image measurements in clinical trials. We will study the effect of each of these sources of variability and will propose a framework for quality assurance.

III. Sources of Image Variability

In this section, we go through the clinical workflow and outline sources of image variability in each stage of the imaging workflow: image acquisition, image post-processing and image interpretation.

A. Image acquisition: Patient cooperation

The patient himself could be a major source of variability in clinical trials. The degree of overall patient cooperation and during image acquisition determines the quality of the produced images. During the scan, there are clinical trial protocols that ask for patient breath hold up to 30 seconds that are otherwise not followed would cause introducing image motion artifacts. In addition, patient mispositioning and external movements could introduce image motion artifacts. In general, patient weight is a factor that should be considered as weight loss or increase may cause the shrinkage or progression of a tumor size misleading the tumor measurements results.

B. Image acquisition: Effect of using different scanners

A multi-center clinical trial may involve Magnetic Renonsance Imaging (MRI) and Computed Tomography (CT) scanners of different manufacturers and software versions. Image quality of MR is more vulnerable to scanner variations than that from CT modality.

Sources of variability in MR include the magnetic field strength, field homogeneity, pulse sequence parameters (TE, TR, flip angle, …etc) and scanner Signal to Noise Ratio (SNR).

A study has shown that follow-up tumor measurements performed using different CT scanner manufacturers with different software versions and different scan protocols could affect the accuracy of automated tumor volumetric size measurement [3].

The modality type may be even changed by some clinical trial protocols along time points in the trial. Modalities used in practice in clinical trials include X-Ray, CT and MR. CT is the preferable modality choice for tumor measurements and follow-up since it is more sensitive than x-ray in showing more

details and discovering new lesions. X-Ray is accepted but not preferable since the pixels represent summations of densities. Although MRI has some complications and variability, it is recommended as a second choice after CT especially when enhanced using a contrast for abdomen. Ultrasound is not reproducible and operator-dependent and hence not recommended for clinical trials. It can only be used to confirm the disappearance of tumors. A special care should be taken when changing the modality for example from CT to MR for tumor assessment which may affect the sharpness of the tumor boundaries.

It is critical for imaging sites to calibrate their imaging scanners using standard phantoms to ensure consistent image quality during the trial.

C. Image acquisition: Effect of slice thickness, kVp and mAs

The slice thickness is one of the parameters that could affect the detectability of new lesions and the measurements of identified lesions at follow-up. Each volume within the slice thickness is scanned and the average attenuation coefficient of each voxel is represented in 2-D cross-sectional image. Depending on how much of the tumor is within the slice thickness, the average density of tumor may be underestimated [4]. The thinner the slice thickness is, the higher will be the patient dose exposure, the higher will be the spatial resolution along the patient axis but the noisier will be the image.

The minimum measurable lesion at baseline should be twice the slice thickness of the scan (e.g. 5 mm slice thickness is required for detecting a minimum lesion size of 10 mm). The minimum size of a lesion at baseline for a given slice thickness ensures that any lesions appearing smaller on subsequent examinations are truly decreasing in size. Therefore, the same slice thickness should be used in all subsequent examinations to avoid variations in the tumor measurements.

A higher tube mA would increase the image signal to noise ratio. As the kVp goes lower at constant mAs, the image contrast becomes higher [5].

D. Image interpretation: Taking image measurements

The DICOM header attribute Pixel Spacing (0028,0030) specifies the physical distance at the patient level between the centers of each pixels in both rows and columns directions. It is used in cross sectional modalities such as CT and MR for performing image measurements by image display software. For other radiographic projection modalities such as CR, MG, DX and XA, the DICOM attribute Imager Pixel Spacing is used which specifies the physical distance at the receptor plane between centers of each pixel. To get a true size measurement, the imager pixel spacing value is divided by the magnification factor attribute (i.e. ratio of source to detector distance to source to patient distance) to correct for magnification effects due to the fact that objects close to the source are more magnified than objects close to the receptor [6]. A DICOM correction proposal (CP586) was introduced to the DICOM standard to resolve the problem of having the attribute pixel spacing present for radiographic modalities [7]. In order to avoid always using this attribute by imaging workstations and ignoring the imager pixel spacing attribute, the CP mentions that if this attribute is present, it should have the same value as the imager pixel spacing if uncorrected. If it is present with a different value, then the image should have been calibrated beforehand either manually by the operator using object of known size and depth within the patient (e.g. burned-in ruler) or automatically using the modality in order to correct for the geometry. A display software should warn the user to calibrate the image before performing image measurements if the pixel spacing attribute is not calibrated. If calibration was not possible, either manually or automatically, the display software should indicate to the user that the measurements are at the level of the detector. If the pixel spacing attribute is only present in radiographic modalities, it should be used but the user should be informed by the calibration type if given or if it is unknown. However, for CT and MR the pixel spacing attribute should be only present and used. If the DICOM images are digitized from films as Secondary Capture (SC) images, then there is no way to get pixel spacing information and hence calibration is not possible unless the films have scale information burned on it.

It should be noted that for ensuring a consistent and authentic image display, the monitors should be calibrated according to standard display function such as the DICOM Grayscale Standard Display Function (GSDF).

Medical guidelines such as RECIST and WHO criteria have been specified to support radiologists ensuring consistent image measurements. Repeated measurements of the lesions should be taken using the same window width and window level values for viewing images. Using different window values may affect the edges of the tumor boundary misleading the interpretation and measurement [4].

A study has shown that CT lung tumor size measurements are inconsistent in clinical trials and that differences in measurements are greatest when the edge of the tumor is irregular or not well-defined [8].

Apparent changes have been reported in CT scans retaken after 15 minutes for the same lung tumor patients where the variability are found to be greatest for small lesions [9].

E. Image post-processing: Effect of lossy image compression

With the increasing amounts of imaging exams volumes, there is a growing need for image compression to reduce storage requirements and allow for a faster network transmission. Lossy image compression is being adopted in clinical practice since they can provide higher compression ratio while achieving acceptable image quality. Comparing images that were lossy compressed against original images that were not compressed may have an effect on the image analysis for performing measurements if the image quality of the lossy compressed images is not suitable for diagnosis.

The evaluation of image compression schemes and their usability for clinical practice has been a research topic since many years. Many studies have focused on determining the degree to which images can be lossy compressed while being visually lossless for performing diagnostic analysis. The compression ratios that are considered acceptable for diagnosis vary according to the type of modality, organ and disease. Quantitative measures such as the compression ratio, Normalized Mean Squared Error (NMSE) (Eq. 1) and Peak Signal to Noise Ratio (PSNR) (Eq. 2) are used for the evaluation of the quality of the reconstructed image $f_A(x, y)$ compared to the original image $f(x, y)$ of size $N \times N$.

$$NMSE = \frac{\sum_{x=0}^{N-1}\sum_{y=0}^{N-1}(f(x,y)-f_A(x,y))^2}{\sum_{x=0}^{N-1}\sum_{y=0}^{N-1}f(x,y)^2} \quad (1)$$

$$PSNR = \frac{20\log(f(x,y)_{\max})}{((\sum_{x=0}^{N-1}\sum_{y=0}^{N-1}(f(x,y)-f_A(x,y))^2)(1/N\times N))^{1/2}} \quad (2)$$

These measures when reported with the lossy compressed images can be used to determine the acceptability of images for diagnosis based on disease type and the best-practice recommendations from other previous research studies results.

IV. Reporting the Measurements

Tumor size can be quantified using one-dimensional, two-dimensional image measurements or 3-D volumetric measurements. According to WHO criteria, the sum of the product of longest lesion diameters of two perpendicular dimensions is reported. According to RECIST criteria, the sum of longest lesions diameters are used as a one-dimensional measurement. However, it has been proven that 2-D measurements are not accurate and reproducible as 3-D measurements. Volumetric measurements are usually performed in an automated fashion using image processing software. The automated volumetric measurements were proved to be more accurate and reproducible compared to manual 2-D image measurements [10].

The tumor measurements are recorded traditionally in tabulated spreadsheets such as Excel sheets using standard templates designed by CDISC. One of the problems of this approach is the constraint that impairs the workflow that the measurements are separate from the images and there is no reliable and reproducible way to review the images along with the associated reported measurements.

DICOM Structured Reporting (SR) has been proposed to record the image measurements so that it becomes possible to import the DICOM SR document to the image archive (PACS) and later review the images with the measurements by radiologists or other experts [11].

If modality is CT or MR
- Find and get the pixel spacing attribute value.
- Find and get the pixel spacing calibration type
- If calibration type not present or if not equal to (GEOMETRY or FIDUCIAL).
  - ➢ Warn the user that the calibration is unknown.

If modality is CR, DX, MG and XA
- Find and get the pixel spacing attribute value.
- Find and get the imager pixel spacing attribute.
- If (pixel spacing == imager pixel spacing)
  - ➢ The image is not calibrated.
  - ➢ Calibrate the image by dividing the imager pixel spacing by the magnification factor attribute.
- Else if (pixel spacing != imager pixel spacing)
  - ➢ Then the image is calibrated
  - ➢ Check the calibration type and warn the user if it is unknown.

Fig. 1. Image Calibration Checking Algorithm

V. Software Tools for Quality Assurance

Open source software tools have been developed for medical image quality assurance before performing image measurements and for reporting those measurements. A software module called "dcmqa" has been implemented for performing quality assurance tasks using the OFFIS DICOM Toolkit (DCMTK) in C++. The tasks include extraction of quantitative measures such as compression ratio, NMSE and PSNR for comparing reconstructed images with original images for evaluating the image quality. The quality assurance tool also will check for basic attributes for image viewing and the DICOM attributes required for performing image measurements. The tool will warn the user if the pixel spacing attribute is not calibrated. The algorithm of image calibration for medical images can be implemented as shown in Fig. 1.

For evaluation, a validated image dataset can be used for performing true-size image measurements of a burned-in ruler with a known size for implementers to test their image measurements tools. The dataset can be downloaded from (http://www.dclunie.com/images/pixelspacingtestimages.zip).

VI. Conclusion

There are many sources of variability that could affect the comparability of image measurements in clinical trials. A special care is needed when interpreting those measurements by clinical trials investigators. Open source software tools for quality assurance were developed to support the clinical trial sponsors and investigators to ensure a consistent image quality during the trial.

Acknowledgment

This work is implemented within the SWABIK project which is funded by the German federal ministry of research and education under the grant number 01 EZ 1023.

References

[1] ICH Expert Working Group, "Guideline for good clinical practice," in International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use, 1996.

[2] RL. Prentice, "Surrogate endpoints in clinical trials: definition and operational criteria," in Stat Medicine, vol. 8, pp. 431-440, 1989.

[3] M. Das and et al., "Accuracy of automated volumetry of pulmonary nodules across different multislice CT scanners," in Euro Radiol, vol. 17, pp. 1979-1984, 2007.

[4] E. Eisenhauer and et al., "New responses evaluation criteria in solid tumors: Revised RECIST guideline (v1.1)," in European J. of Cancer, vol. 45, pp. 228-247, 2009.

[5] European Commission, "European guidelines on quality criteria for diagnostic radiographic images," in EUR 162260, 1999.

[6] D. Clunie, "On the matter of size and distance measurements in digital mammography," white paper, 2010.

[7] DICOM Standards Committee, Digital Imaging and Communications in Medicine (DICOM) Correction Proposal 586, "Pixel spacing and calibration in projection radiography," National Electrical Manufacturers Association, 2006.

[8] J. Erasmus, “Interobserver and intraobserver variability in measurement of non-small cell carcinoma lung lesion: Implications for assessment of tumor response,” in J. Clin. Oncol, vol. 21, pp. 2574-2582, 2003.

[9] G. Oxnard and et al., “Variability of lung tumor measurements on repeat computed tomography scans taken within 15 minutes,” J. Clin. Ongol, vol. 29, pp. 3114-3119, 2011.

[10] K Marten, F Auer, S Schmidt, G Kohl, EJ Rummeny, C Engelke, “Inadequacy of manual measurements compared to automated CT volumetry in assessment of treatment response of pulmonary metastases using RECIST criteria,” in J. Eur Radiol vol. 6, pp. 781-790, 2006.

[11] D Clunie, “DICOM structured reporting and cancer clinical trials results,” in J. Cancer Informatics, vol. 4, pp. 33-56, 2007.