

A Tool for Collection of Region Based Data from Uterine Cervix Images for Correlation of Visual and Clinical Variables Related to Cervical Neoplasia

Jose Jeronimo, Mark Schiffman

*Hormonal and Reproductive Epidemiology Branch, Division of Cancer Epidemiology and Genetics (DCEG), National Cancer Institute, National Institutes of Health, Department of Health and Human Services, Bethesda, MD
guibovij@mail.nih.gov*

L. Rodney Long, Leif Neve, Sameer Antani

*Communications Engineering Branch, Lister Hill National Center for Biomedical Communications, National Library of Medicine, National Institutes of Health, Department of Health and Human Services, Bethesda, MD
{rlong, lneve,santani}@mail.nih.gov*

Abstract

The National Cancer Institute (NCI) is collaborating with the National Library of Medicine (NLM) to create a database of digitized images of the uterine cervix for research, training, and education. The database of 100,000 images collected in NCI projects will be Web-accessible and will contain not only the digitized images, but also clinical, longitudinal data relevant to the detection of uterine cervix cancer. The database will also contain spatial data in the form of expert-marked boundary regions of anatomy and tissue regions of high interest. The collection of this spatial data is enabled by customized software, the Boundary Marking Tool (BMT). The BMT, its significance, and its use in preliminary research studies, are described in this paper.

1. Background: source of data and motivation for data collection

Cervical cancer is the second most common cancer affecting women worldwide [1], and the first in many developing countries [2]. Every year there are 400,000 new cases of invasive cervical cancer in the world, and 15,000 of them occur in the United States. This neoplasia is closely related to chronic infection by anogenital types of Human Papillomavirus (HPV), which are sexually transmitted [3]. There are more than 40 different anogenital types of HPV, but only about 15 of them are considered oncogenic because of their relation with pre-invasive lesions and cancer.

The Division of Cancer Epidemiology and Genetics (DCEG) of the National Cancer Institute has been involved in two of the major projects in cervical cancer worldwide, the Guanacaste and ALTS projects, designed to study the natural history of the HPV infection and cervical neoplasia. The Guanacaste Project is a population-based natural history study of HPV and cervical neoplasia in a rural area of Costa Rica, where the rates of cervical cancer are perennially high. Approximately ten thousand (10,000) women were enrolled and followed for up to seven years. State-of-the-art visual, cytological, and molecular screening tests were used

to examine the origins of cervical precancer/cancer and to explore which factors make a geographic region "high risk" [4]. The second project is known as ASCUS LSIL Triage Study (ALTS), and was developed in four geographical areas of the United States, enrolling 5060 volunteer patients with abnormal reference Pap smear of ASCUS (equivocal) or LSIL (mildly abnormal) [5].

The visual data collected during these projects are represented by cervicography, digital images taken at colposcopy (Denvu), cytology (Pap smears) and histology (biopsies). Cervicography, first described by Stafl in 1981 [6], refers to the acquisition of 35 mm photographs of the cervix taken using a conventional camera with a 100 mm macro lens. The camera has a fixed focus to preserve a constant distance between the camera and the cervix; a property allows us to have comparable pictures of the cervixes of all patients, and to perform a measurement of areas of the cervix [7]. During the image acquisition process the photographer manually moves the camera back and forth to get the image in focus. Additionally, the camera has a ring flash for better illumination of the cervix. Immediately before the pictures are taken, the cervix is washed with 5% acetic acid for one minute. A pair of pictures is taken on each visit to be sure that at least one of them is good enough for evaluation. Later, an expert in cervicographic evaluation projects those pictures onto a screen to get a magnified image of the cervix.

HPV infected squamous epithelium often turns white when is exposed to 5% acetic acid (vinegar) which is called "acetowhite" epithelium. Colposcopists have used this phenomenon for decades to study the cervix and to guide biopsies. During the colposcopic exam, the evaluators look for some characteristics of the infected or precancerous epithelium, such as whitening, vascular pattern, margins, etc. These findings enable estimation of the stage of the disease [8]. Theoretically, there should be a logical sequence from visual changes indicative of HPV infection to changes implying precancer, but in practice the distinction is difficult to make reproducibly.

During the ALTS and Guanacaste projects, approximately 100,000 cervigrams were taken from patients with invasive cancer or intraepithelial lesions; women who were healthy at enrollment but developed disease during the follow-up; healthy women who never developed pathological changes in the cervix, etc. This huge amount of visual information provides a great opportunity for studying the uterine cervix changes related or not related to HPV infection and/or cervical neoplasia; but at the same time, it presents a challenge to manage thousands of pictures, and is in a practical sense impossible to do through the direct evaluation of the conventional 35 mm slides. It has been necessary to develop strategies and software tools for the digitization of the slides; and the compression, management, and study of the resulting digitized pictures.

2. Overview of the software system for uterine cervix data archiving, research, and dissemination

The system being built for archiving, research and dissemination of the digitized Guanacaste images has two components: first, the Boundary Marking Tool (BMT), which is intended primarily for the collection of boundary region data in the images; and, second, the Multimedia Database Tool (MDT), which supports Web dissemination of the images, the related clinical textual data, and the boundaries collected with the BMT. The BMT is designed to operate in a stand-alone mode for data collection (mainly for collecting boundary region data). BMT display functionality is being incorporated into the MDT to allow the display of BMT-acquired boundaries. The BMT stores its outputs in a MySQL database, which is then

read (under control of a user's query) by the MDT. BMT development is complete; MDT development is ongoing, but an early prototype exists for demonstration purposes.

3. The Boundary Marking Tool

The BMT allows the collection of a detailed set of data relevant to the evaluation of uterine cervix images, including adequacy of the image for visual evaluation, visual diagnosis,

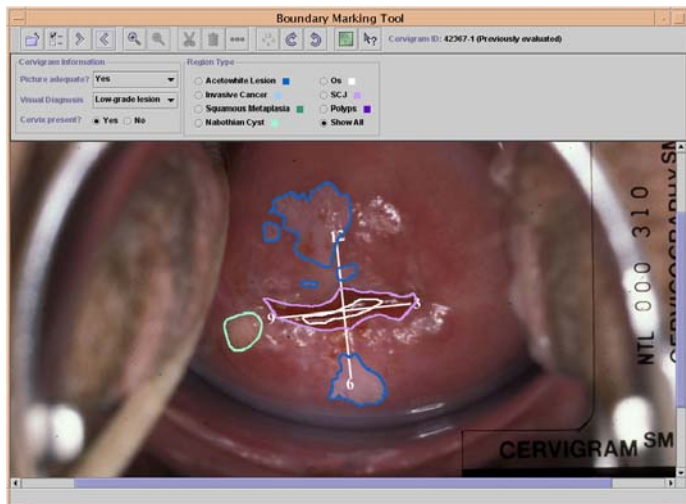


Figure 1. Boundary Marking Tool with color-coded regions and 3/6/9/12 landmarks shown.

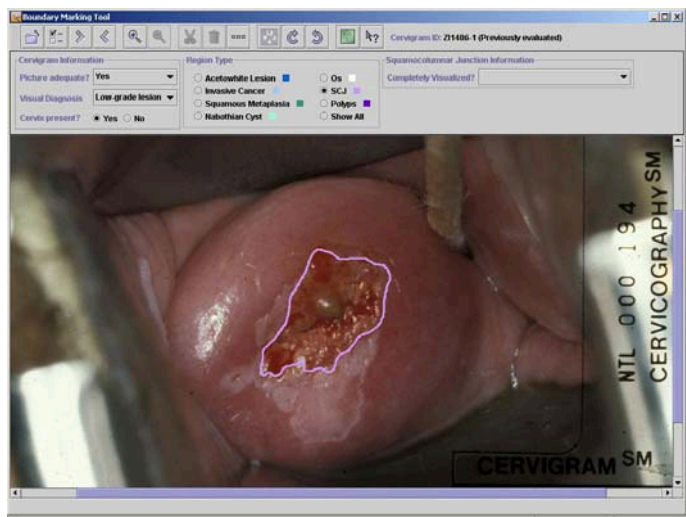


Figure 2. BMT with boundary around the Squamo-columnar junction.

the presence or absence of visual “tiling” (mosaicism) is recorded, along with a classification of the mosaicism as coarse or fine. In addition, the margins and color of the region are classified using a standard four-level index. All BMT outputs are saved as records in a central MySQL database that resides on a server at NLM; these outputs include the spatial boundary data, which

presence of the cervix, and the marking of a set of “3/6/9/12” orientation landmarks, as well as anatomical regions corresponding to areas of acetowhite epithelium, suspicious or presence of invasive cancer, areas with squamous metaplasia (a benign change of the cervix easy to confound with pre-malignant epithelium), cysts, the cervical os, polyps, and the squamous-columnar junction (SCJ) that is the area where the two epithelia of the cervix converge and more than 90% of the cervical cancers arise (Figure 1). Where appropriate, the expert is provided

with a pull-down menu of choices. For example, for “visual diagnosis”, the pull-down menu includes the choices of normal, cervicitis, metaplasia, condyloma, low-grade lesion, high-grade lesion, and invasive cancer.

When the expert draws boundaries on the image the boundaries are shown in color-coded form. Capability is provided to indicate that a boundary is partially obscured and to record an assumed boundary path for the obscured area. For particular regions, the expert may display a detailed view that allows sub-classification of the region contents. For example, for acetowhite regions,

is recorded as a set of (x,y) pixel coordinates in a standard image coordinate system. The BMT is a cross-platform Java application that runs on the user's desktop and loads digitized cervigrams from local storage (e.g., DVD) while saving results to the central MySQL database by means of the MySQL Connector/J package which implements Sun's Java Database Connectivity (JDBC) API.

4. Preliminary biomedical research use of the Boundary Marking Tool

The BMT has been already used in a study developed to explore HPV tropisms for the glandular or squamous epitheliums of the cervix. Based on the results of a previous study among hysterectomized women that demonstrated a high prevalence of some types of HPV infecting the squamous epithelium of the vagina, we hypothesized that some types of HPV virus would be more likely to be found in the squamous epithelium, while some other types could infect the glandular epithelium of the cervix. Pictures from 1016 patients from the Guanacaste project were reviewed to identify the junction between those epitheliums, and then a boundary was drawn around it. The area obtained, called ectopy, represents the amount of glandular epithelium exposed on the ectocervix (Figure 2). The preliminary analysis of the results of this study shows that there is a relationship between the amount of ectopy and the HPV detected. The evaluation will be extended to a second group of 800 patients infected with HPV during the follow-up to confirm our findings.

Currently, the Division of Cancer Epidemiology and Genetics of the NCI has initiated a partnership with the American Society of Colposcopy and Cervical Pathology to use the BMT for developing research studies on HPV infection and cervical cancer. These studies will be mainly focused on the evaluation of the changes produced by the HPV on the uterine cervix, to determine the chain of visual events that lead some patients to develop cancer and other patients to clear the virus. Also we will explore the differences in visual manifestations of the several kinds of HPV. Additionally, because there are thousands of pictures of healthy women, it is possible to study the characteristics of the normal cervix. These results can later be used as a reference to compare with pictures from patients who developed cervical pathology.

5. Research use of the Boundary Marking Tool for management of large volumes of images

Data marked through the BMT provide a rich resource that can be used in the management of the large volume of digitized 35mm slides of the uterine cervix. The MDT is proposed as a tool for Web dissemination of the images, the related clinical textual data, and the boundaries collected with the BMT. These would be searchable by text descriptors for various data fields in a manner similar to that adopted by WebMIRS¹. This approach must be augmented by image management techniques, however, for enabling research on expert marked boundaries. One approach is to adopt image content analysis methods for studying the image properties in the regions marked by the experts and to use the computed features for indexing the images for enabling efficient retrieval by content description. Research in this has been typically referred as content-based image retrieval (CBIR).

Color and texture are expected to be the predominant features used in technical work to index the images by image content. Analysis of expert-marked regions will enable correlation between similar image structures and also support a variety of classifications that may be of

¹ WebMIRS: Web-based Medical Information Retrieval System, see <http://archive.nlm.nih.gov/proj/webmirs>

interest to the researcher. In the education and training domain, such techniques can be used to assist in retrieval of images that exhibit a particular visual feature. A side benefit of this indexing initiative is better image management through the development of a database indexed on image features.

6. Conclusion

The BMT is a valuable tool that facilitates the effective use of medical images for research and training purposes, by enabling the collection, archiving to database, and retrieval of text and spatial data for the study of cervical neoplasia in uterine cervix images. Currently, DCEG has a list of more than 12 studies to be developed with gynecologist experts in cervical cancer working at geographically dispersed sites in the United States. These experts will use the BMT through the World Wide Web. The knowledge obtained will allow us to understand more clearly the natural history of HPV infection and cervical cancer.

References

- [1] Parkin DM, Bray FI, Devesa SS. Cancer burden in the year 2000. The global picture. *Eur J Cancer* 2001; 37 Suppl 8:S4-66.
- [2] Eluf-Neto J, Nascimento CM. Cervical cancer in Latin America. *Semin Oncol* 2001; 28(2):188-197.
- [3] Bosch FX, Lorincz A, Munoz N, Meijer CJ, Shah KV. The causal relation between human papillomavirus and cervical cancer. *J Clin Pathol* 2002; 55(4):244-265.
- [4] Herrero R, Schiffman MH, Bratti C, Hildesheim A, Balmaceda I, Sherman ME et al. Design and methods of a population-based natural history study of cervical neoplasia in a rural province of Costa Rica: the Guanacaste Project. *Rev Panam Salud Publica* 1997; 1(5):362-375.
- [5] Schiffman M, Adriaenza ME. ASCUS-LSIL Triage Study. Design, methods and characteristics of trial participants. *Acta Cytol* 2000; 44(5):726-742.
- [6] Staffl A. Cervicography: a new method for cervical cancer detection. *Am J Obstet Gynecol* 1981; 139(7):815-825.
- [7] Jacobson DL, Peralta L, Farmer M, Graham NM, Wright TC, Zenilman J. Cervical ectopy and the transformation zone measured by computerized planimetry in adolescents. *Int J Gynaecol Obstet* 1999; 66(1):7-17.
- [8] Reid R, Scalzi P. Genital warts and cervical cancer. VII. An improved colposcopic index for differentiating benign papillomaviral infections from high-grade cervical intraepithelial neoplasia. *Am J Obstet Gynecol* 1985; 153(6):611-618.