Published in final edited form as:

AIDS. 2022 March 01; 36(3): 409–413. doi:10.1097/QAD.0000000000003123.

Using Computer-Assisted Content Analysis to Advance Anal Dysplasia Natural History Research: A Validation Study

Edward R. CACHAY[1], **Farnaz HASTEH**[2], **Wm. Christopher MATHEWS**[3]

[1]Department of Medicine, Division of Infectious Diseases and Global Public Health, UC San Diego

[2]Department of Pathology UC San Diego

[3]Department of Medicine, UC San Diego

Abstract

Objective: Our study aim was to validate use of computer-aided narrative content analysis in the extraction of standard diagnostic categories using an archived cytology database that included individually overread reference classification.

Design: Retrospective analysis of narrative anal cytology results collected on HIV-infected patients at the UCSD between January and December 2001.

Methods: We used computer-assisted content analysis extraction methodology using Wordstat 8.0 (Provalis Research) that operated using a classification dictionary that we developed for the following diagnostic categories: NAMC, ASCUS, LSIL, HSIL. We compared its accuracy to a physician overread manually extracted methods that classified each report into the most severe diagnostic category referenced in the narrative report. Agreement between content analysis mapped diagnostic categories and the reference category was evaluated using kappa agreement.

Results: During 2001, 901 patients underwent 997 anal cytological examinations as routine screening. By reference diagnostic category: 54 (5.4%) were unsatisfactory, 460 (46.1%) were NAMC, 291 (29.2%) were ASCUS, 131 (13.1%) were LSIL, and 61 (6.1%) were HSIL. Computer-aided content analysis extracted a single diagnosis from each report in 963 (96.2%) cases and two diagnoses in 38 (3.8%) cases. The Kappa agreement was 0.96 (0.019 s.e.). There were 29 cases classified ASCUS by reference category but LSIL by adjudicated content analysis. A focused review indicated that the over reader assigned reference category was in error.

Conclusions: Computer-aided narrative content analysis of anal cytology results yielded accurate and time-efficient classification into meaningful diagnostic categories that can be used to evaluate screening programs and modeling natural history.

Corresponding author: Edward Cachay, Please address correspondence to: Edward Cachay MD, MAS, 200 West Arbor Drive, San Diego, CA 92103-8186, Ph: 619-543-3882, Fax: 619-543-3178, ecachay@health.ucsd.edu.

Author contributions:
Edward Cachay: study concept and design, analysis interpretation, drafting of the manuscript, study supervision, approved the final submission
Faranaz Hasteh: Study design feedback critical revision of the manuscript, approved the final submission
Wm. Christopher Mathews: Study design, statistical analysis, and interpretation of data, critical revision of the manuscript, writing and approved the final submission.

Keywords

Content analysis; Anal cytology; anal dysplasia; HIV

Background:

Persons living with HIV (PWH) are at increased risk of anogenital human papillomavirus (HPV) infection and related cancers [1,2]. It is estimated that HPV-related anal cancers will account for a significant fraction of cancers in PWH by 2030 [3]. Consequently, many centers have implemented anal screening programs incorporating anal cytology tests followed by high-resolution anoscopy (HRA) and anal punch- biopsy [4].

We have previously shown that 1 of 133 PWH with an anal cytology result of high grade squamous intraepithelial lesion (HSIL), the immediate precursor of anal cancer, will develop invasive cancer annually [5]. To better understand the natural history of progression to anal cancer, we need to monitor systematically many individuals at risk. Potentially, sizeable integrated longitudinal data sets with well-characterized clinical information from anal screening programs can address this need. Many cancer screening programs work with existing electronic medical records (EMR) [6,7]. Coding systems such as Logical Observation Identifiers Names and Codes (LOINC), Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT), and The International Classification of Diseases 10th edition (ICD10) have codes for cytology diagnosis categories [8–10]. However, a major logistical challenge to longitudinally evaluating anal cancer screening program outcomes is the inconsistent use of existing codes cytology diagnosis categories that would facilitate quantitive analysis and modeling of coded severity categories. Manual data abstraction and coding of the narrative result field of cytology reports in large data sets is labor-intensive and a significant barrier to pooling data sets across screening programs.

In recent years, there has been a focus on using natural language processing and computer-aided content analysis to overcome the limitations of narrative content available in EMRs [11]. Such approaches might facilitate health services and epidemiological research concerning natural history, response to treatment, and screening program outcomes. Therefore, our study aim was to validate the use of computer-aided narrative content analysis in the extraction of standard diagnostic categories using an archived cytology database that included individually overread reference classification.

Methods:

We conducted a retrospective analysis of a data set of archived narrative anal cytology results collected in enrolled adult (≥18 years) PWH attending an HIV primary care appointment at the University of California, San Diego (UCSD) Owen Clinic during the first year of our anal screening program implementation in 2001. This data set was chosen because it included a reference standard cytology diagnosis categories assigned by an individual physician overreader who classified each report into the most severe diagnostic category referenced in the narrative report. According to our standard of care, PWH with any abnormal anal cytology result were referred to our collocated anal screening clinic, where patients underwent an HRA evaluation. All participants signed written consent before study enrollment. The study protocol was approved by the UCSD Institutional Human Research Protection Program (Project no. 150 186).

We use Wordstat 8.0(Provalis Research) as the text analytical tool for the content analysis extraction[12]. First, we developed a classification dictionary for the content analysis for the following diagnostic anal cytology categories: unsatisfactory, no atypical or malignant cells (NAMC), atypical cells of uncertain significance (ASCUS), low grade squamous intraepithelial lesion (LSIL), high grade squamous intraepithelial lesion (HSIL). Noteworthy, the category “atypical squamous cells, cannot exclude high grade,” introduced in the Bethesda 2001 revision, had not yet been implemented.

Agreement between content analysis mapped diagnostic categories and the reference category was evaluated using kappa agreement with 95% bootstrapped confidence intervals (C.I.). Statistical analysis was conducted using Stata Version 16.1.

Results

Between January 2001 and December 2001, 901 patients underwent 997 anal cytological examinations as part of routine screening. Table 1 presents the Wordstat 8.0 classification dictionary to map the narrative cytology result elements to specific diagnosis categories.By the reference diagnostic category, 54 (5.4%) of anal cytology results were unsatisfactory, 460 (46.1%) were NAMC, 291 (29.2%) were ASCUS, 131 (13.1%) were LSIL, and 61 (6.1%) were HSIL. The computer-aided content analysis extracted a single diagnosis from each report in 963 (96.2%) of the cases and two diagnoses in 38 (3.8%) cases. Table 2 presents content analysis adjudicated most severe diagnostic category by physician overreader assigned reference category. The kappa agreement was 0.96 (95% C.I.: 0.940 – 0.972). There were 29 cases in which anal cytology results were classified as ASCUS by reference category but LSIL by adjudicated content analysis. These 29 cases were mapped to 2 diagnostic categories by content analysis (ASCUS and LSIL). A focused review of the narrative reports in each of the 29 cases indicated a coding error in the reference category, in that the historical reviewer should have assigned the severest diagnostic category (LSIL).

Discussion

This study validates the use of computer-aided content analysis for efficient extraction of coded diagnosis categories from EMR anal cytology narrative reports. Agreement between content analysis assigned anal cytology diagnostic categories and those assigned by an individual physician overreader was high. In fact, it may be that computer-aided assignment of categories may be more accurate than labor-intensive but fallible individual record review, particularly when more than one codable diagnosis is included in the narrative cytology report. Our results suggest that such an approach could allow incorporation of routinely collected but heretofore logistically inaccessible narrative EMR data into existing repositories of clinical and laboratory data, thereby enhancing their usefulness for epidemiological evaluation of anal cancer screening programs

HPV-related anal cancer is a relatively slow disease, and clinical trials are limited to evaluating the natural history of anal cancer as the primary outcome because most of them follow individuals at risk for five years or less [13]. Further, the large number of participants required makes funding sustainability very challenging. To understand its natural history more accurately, we need big data analytics approaches leveraging enriched clinical and laboratory data from anal cancer screening programs [14].

Some limitations must be noted. We used an archived anal cytology cohort assembled before 2001 Bethesda System for cytology reporting had been implemented. Thus we could not evaluate the performance of our category assignment dictionary in accurate recognition of the diagnosis category *atypical squamous cells, cannot rule out HSIL (ASC-H)* [15]. Our previous work, however, has suggested the ASC-H is associated with a similar progression probability as HSIL and may reasonably be combined with HSIL in modeling studies [16]. Additionally, we did not validate text report analysis of histopathology results from HRA-directed biopsies. We plan to address this in a second step. Yet, our intention is to share an efficient and accurate strategy with investigators in the field that may allow access to critical clinical information stored in narrative text. This approach can facilitate clinical, epidemiological, and translational studies addressing the natural history of anal neoplasia and its precursors in populations at risk for anal cancer, including PWH and other immunosuppressed individuals at risk, too, such as persons who received organ transplantation.

Funding Support:

This work was supported in part by The University of California San Diego Center for AIDS Research [AI036214], and the Pacific AIDS Education and Training Center (PAETC). The funders had no role in study design, decision to publish, or preparation of the manuscript.

Bibliography:

1. Mahale P, Engels EA, Coghill AE, Kahn AR, Shiels MS. Cancer Risk in Older Persons Living With Human Immunodeficiency Virus Infection in the United States. Clin Infect Dis. 2018;67:50–57. [PubMed: 29325033]
2. Colón-López V, Shiels MS, Machin M, Ortiz AP, Strickler H, Castle PE, Pfeiffer RM, Engels EA. Anal Cancer Risk Among People With HIV Infection in the United States. J Clin Oncol. 2018;36:68–75. [PubMed: 29140774]
3. Shiels MS, Islam JY, Rosenberg PS, Hall HI, Jacobson E, Engels EA. Projected Cancer Incidence Rates and Burden of Incident Cancer Cases in HIV-Infected Adults in the United States Through 2030. Ann Intern Med. 2018;168(12):866–873. [PubMed: 29801099]
4. Mathews C, Caperna J, Cachay ER, Cosman B. Early impact and performance characteristics of an established anal dysplasia screening program: program evaluation considerations. Open AIDS J. 2007;1:11–20. [PubMed: 18776956]
5. Cachay E, Agmas W, Mathews C. Five-year cumulative incidence of invasive anal cancer among HIV-infected patients according to baseline anal cytology results: an inception cohort analysis. HIV Med. 2015;16:191–5. [PubMed: 25197003]
6. Cowburn S, Carlson MJ, Lapidus JA, DeVoe JE. The association between insurance status and cervical cancer screening in community health centers: exploring the potential of electronic health records for population-level surveillance, 2008–2010. Prev Chronic Dis. 2013;10:E173. [PubMed: 24157076]

7. Petrik AF, Green BB, Vollmer WM, Le T, Bachman B, Keast E, Rivelli J, Coronado GD. The validation of electronic health records in accurately identifying patients eligible for colorectal cancer screening in safety net clinics. Fam Pract. 2016;33:639–643 [PubMed: 27471224]
8. National Cancer Institute, enterprise Vocabulary Services. LOINC codes for abnormal anal cytology Available at: https://nciterms.nci.nih.gov/ncitbrowser/ConceptReport.jsp?dictionary=LOINC&code=77653-4&ns=LOINC. Accessed 25 march 2021
9. Centers for Diseases Control and Prevention, Public Health Information Network Vocabulary Access and Distribution System. SNOMED CT codes for abnormal anal Papanicolaou smear. Available at https://phinvads.cdc.gov/vads/ViewCodeSystemConcept.action?oid=2.16.840.1.113883.6.96&code=439855007 accessed 29 March 29 2021
10. ICD 10 codes for abnormal anal cytology, available at: https://icd.codes/icd10cm/R85613
11. Galetsi P, Katsaliaki K. Big data analytics in health: an overview and bibliometric study of research activity. Health Info Libr J. 2020;37:5–25
12. Available at https://provalisresearch.com/resources/tutorials/wordstat-8-new-features/, accessed 14 April 2021.
13. Poynten IM, Jin F, Roberts JM, Templeton DJ, Law C, Cornall AM, Molano M, Machalek DA, Carr A, Farnsworth A, Tabrizi S, Phillips S, Fairley CK, Garland SM, Hillman RJ, Grulich AE. The Natural History of Anal High-grade Squamous Intraepithelial Lesions in Gay and Bisexual Men. Clin Infect Dis. 2021;72:853–861. [PubMed: 32342984]
14. Borges do Nascimento IJ, Marcolino MS, Abdulazeem HM, Weerasekara I, Azzopardi-Muscat N, Gonçalves MA, Novillo-Ortiz D. Impact of Big Data Analytics on People's Health: Overview of Systematic Reviews and Recommendations for Future Studies. J Med Internet Res. 2021 ;23:e27275 [PubMed: 33847586]
15. Solomon Diane, and Nayar Ritu. The Bethesda System for Reporting Cervical Cytology : Definitions, Criteria, and Explanatory Notes. 2nd ed. New York: Springer, 2004.
16. Mathews WC, Agmas W, Cachay ER, Cosman BC, Jackson C. Natural history of anal dysplasia in an HIV-infected clinical care cohort: estimates using multi-state Markov modeling. PLoS One. 2014 ;9:e104116 [PubMed: 25101757]

Table 1.

Data coded dictionary for the content analysis for the different anal cytology categories

CODED CATEGORY	NARRATIVE MAPPERS
UNSATISFACTORY	UNSATISFACTORY
UNSATISFACTORY	UNSAT*
ASCUS	@NOT_ASC-H [ATYPICAL_SQUAMOUS_CELLS NOT BEFORE CANNOT_EXCLUDE /A /D10]
ASCUS	ASCUS
ASCUS	ATYPIA
ASCUS	ATYPICAL_SQUAMOUS_CELLS_OF_UNCERTAIN_SIGNIFICANCE
LSIL	AIN-1
LSIL	AIN-1
LSIL	AIN1
LSIL	ATYPICAL_SQUAMOUS_CELLS_OF_UNDETERMINED_SIGNIFICANCE_LOW_GRADE SQUAMOUS_INTRAEPITHELIAL_LESION_
LSIL	LESION
LSIL	LGSIL
LSIL	LOW-GRADE_SQUAMOUS_INTRAEPITHELIAL_LESION
LSIL	LOW_GRADE*
LSIL	LSIL
LSIL	LOW_GRADE_SQUAMOUS_INTRAEPITHELIAL_LESION
LSIL	MILD_DYSPLASIA
HSIL	AIN-2
HSIL	AIN-3
HSIL	AIN2
HSIL	AIN3
HSIL	AIN_2-3
HSIL	HGSIL
HSIL	HSIL
HSIL	HIGH_GRADE_SQUAMOUS_INTRAEPITHELIAL_LESION
HSIL	MODERATE_DYSPLASIA
HSIL	SEVERE_DYSPLASIA
ASC_H	ASC-H
ASC_H	ASC/H
ASC_H	ATYPICAL_SQUAMOUS_CELLS-CANNOT_EXCLUDE
ASC_H	ATYPICAL_SQUAMOUS_CELLS_-_CANNOT_EXCLUDE
ASC_H	ATYPICAL_SQUAMOUS_CELLS_CANNOT_EXCLUDE_A_HIGH_GRADE_SQUAMOUS-_INTRAEPITHELIAL_LESION_
ASC_H	ATYPICAL_SQUAMOUS_CELLS_CANNOT_EXCLUDE_A_HIGH_GRADE_SQUAMOUS-_INTRAEPITHELIAL_LESION_
ASC_H	ATYPICAL_SQUAMOUS_CELLS_CANNOT_EXCLUDE_A_HIGH_GRADE_SQUAMOUS _INTRAEPITHELIAL_LESION_

CODED CATEGORY	NARRATIVE MAPPERS
ASC_H	ATYPICAL_SQUAMOUS_CELLS_CAN’T_RULE_OUT
ASC_H	CAN’T_EXCLUDE
ASC_H	CAN’T_R/O
ASC_H	CANNOT_EXCLUDE
ASC_H	CANNOT_EXCLUDE_A
ASC_H	SQUAMOUS_INTRAEPITHELIAL_LESION_
ASC_H	CANNOT_RULE_OUT
SCC	SCC
SCC	SCARCINOMA
SCC	SQUAMOUS_CELL_CARCINOMA
NAMC	NEGATIVE_FOR_INTRAEPITHELIAL_LESION
NAMC	NEGATIVE_FOR_MALIGNANT_CELLS
NAMC	NO_ATYPICAL [*]
NAMC	NO_ATYPICAL_OR_MALIGNANT_CELLS
NAMC	NAMC
SIL_NOS	SQUAMOUS_INTRAEPITHELIAL_LESION_

[*] It is a wildcard for any text that follows it.

@NOT is a negation indicator meaning EXCLUDE ASC-H from being classified as ASC-US

Table 2:

Adjudicated Diagnostic Category by Reference Diagnostic Category

Reference Diagnostic Category	Adjudicated Diagnostic Category					
	ASCUS	HSIL	LSIL	NAMC	Unsatisfactory	Total
ASCUS	262	0	29	0	0	291
HSIL	0	61	0	0	0	61
LSIL	0	0	131	0	0	131
NAMC	0	0	0	460	0	460
Unsatisfactory	0	0	0	0	54	54
Total	262	61	160	460	54	997

kappa agreement was 0.96, (95% C.I.: 0.940 – 0.972).