Journal of Forensic and Legal Medicine 52 (2017) 172–180

Contents lists available at ScienceDirect

Journal of Forensic and Legal Medicine

journal homepage: www.elsevier.com/locate/jflm

# Intra- and inter-rater agreement of the Genital Injury Severity Scale

Dana L. Kelly [a, *], Hillary J. Larkin [a], Lauri A. Paolinetti [a, b]

[a] Alameda Health System, Highland, Department of Emergency Medicine, 1411 E. 31st St, Oakland, CA 94602 USA
[b] Samuel Merritt University, 3100 Summit Street, Oakland, CA 94609 USA

ARTICLE INFO

*Article history:*
Received 13 March 2017
Received in revised form
11 September 2017
Accepted 15 September 2017
Available online 19 September 2017

*Keywords:*
Genital Injury Severity Scale
Inter-rater agreement
Sexual assault
Consensual intercourse
Reliability

ABSTRACT

*Introduction:* The Genital Injury Severity Scale (GISS) is a clinimetrically-tested tool in use for quantifying and qualifying external genital injury after sexual intercourse.
*Purpose:* To evaluate inter- and intra-rater agreement of the GISS amongst examiner/raters in an urban, ethnically diverse, emergency department based sexual assault center.
*Methods:* The study was conducted in three phases. Six examiners with various years of experience rated their own cases and each others' cases greater than one year after the initial exam. They rated the photographs and documentation of each case at least one year apart. Another six raters utilized a combination of the photos and documentation simultaneously from the same cases. The evaluation method was the completion of the GISS for each phase.
*Results:* Based on the experience level of the rater, the differences in overall agreement were not significant. Strength of agreement was highest with the combination of photos and documentation with W ranging from 0.60501 (substantial) to 0.91056 (almost perfect). The GISS variables with the highest level of agreement were tissue break type and toluidine blue uptake type, both with photo evaluation alone and combination of documentation and photos (W = 0.72051 and 0.74599, respectively).
*Conclusion:* The Genital Injury Severity Scale is a reliable tool to quantify and qualify the severity of external genital injury when used to evaluate a combination of photos and documentation utilizing midlevel providers trained as sexual assault forensic examiners with various years of experience.


## 1. Introduction

Sexual assault (SA) continues to be a worldwide problem. According to the National Crime Victimization Survey[1] 284,350 persons in the US aged 12 and older reported having been sexually assaulted in 2014, excluding sexual assaults that resulted in serious injury. This translates into 32.4 victimizations per hour or approximately one every two minutes. Only 33.6% of these incidents were reported to law enforcement agencies.

Following jurisdictional procedure, many victims of sexual assault are offered collection of forensic evidence. These forensic exams are performed worldwide by practitioners with varying levels and types of experience. Regardless of the experience profile of the examiner, there is a minimum expectation that a determination be made as to whether the examination findings are consistent with the history given by the victim. To do this, as well as to collect appropriate evidence, examiners must have specialized forensic training.[2] They are also required to possess knowledge of the current evidence based literature, reporting analyses of data collected regarding physical findings after sexual assault. It is the knowledge assimilated from this literature, combined with the examiners' experience that facilitates their interpretation of the exam findings, both in the course of the forensic exam and during expert testimony.

The Federal Rules of Evidence, rules 702 and 703 stipulate that scientific methods used to guide expert testimony should be valid. They, therefore, should be scientifically tested, i.e., have a known rate of error, have established standards, and be subjected to peer review and scientific rigor.[3]

## 2. Literature review

To date, the SA literature lacks an instrument for multidimensional quantification and qualification of genital injury after SA.

* Corresponding author.
*E-mail addresses:* dkellypac@gmail.com (D.L. Kelly), hlarkin@alamedahealthsystem.org (H.J. Larkin), lpaolinetti@samuelmerritt.edu (L.A. Paolinetti).

https://doi.org/10.1016/j.jflm.2017.09.011
1752-928X/

Recent research on physical findings after SA has focused primarily on the mere presence or absence of a pre-defined type or prevalence of injury as its outcome variable.[4–6] Many SA studies have, to some degree, determined the inter-rater agreement among a variety of raters, utilizing various injury patterns and detection methods as variables.[5–10] There are four studies that looked specifically at forensic examiners' agreement among various categories of genital injuries utilizing photographs taken during the examinations and photographic characteristics. Ernst et al.[11] reported a "high degree" of agreement regarding photographic quality utilizing their own previously-developed Photo Documentation Image Quality Scoring System (PDIQSS ©). This study did not evaluate the content of the photographs. Sachs et al.[12] evaluated the reliability of SA examiners to detect genital injury from digital macro computer images. Their study resulted in an overall agreement of 82% (kappa 0.57), utilizing the variables of "perfect", "moderate", and "poor" agreement in evaluating the findings of tears, ecchymosis, and abrasions. In Astrup et al.[13] various raters evaluated colposcopic photographs of consensual intercourse and SA subjects aged over 14 years to identify lacerations, abrasions, and contusions/hematomas/bruises. The overall agreement strength was moderate (kappa 0.41). Sommers et al.[14] analyzed inter- and intra-rater reliability of specially trained novice raters as compared to color analysis experts utilizing specific color spectrum characteristics of colposcopic photographs. They concluded that a standard protocol utilizing multiple, trained analysts be employed for further digital image color analysis.

## 3. The GISS development

The GISS was first introduced in 2012 in the form of a pilot study[15] that was a primary component of validity testing in the scale's derivation.[16] It was developed for use in quantifying and qualifying external genital injury in the adolescent and adult female after both consensual and non-consensual sexual intercourse. It has been validated as a tool to assist in distinguishing sexual assault (SA) from consensual intercourse (CI). For each study, the cohorts stemmed from an ethnically diverse urban population in an inter-city emergency department in the US utilizing midlevel practitioner (MLP: physician assistants and nurse practitioners) forensic examiners.

The clinical model of the GISS, shown in Fig. 1, has been determined, through principal components analysis and internal consistency testing, to be a clinical model that yields the most information utilizing the least number of variables. Its design is suited for determining a multidimensional depiction of external genital injury in women reporting consensual or non-consensual sexual intercourse.

## 4. Utilization of the Genital Injury Severity Scale

An examiner completes the GISS immediately following the external genital examination with utilization of magnification and toluidine blue dye. The appropriate severity (column) of each individual finding (row) is selected based on the results of the examination. The examiner identifies the cell representative of the most severe observed finding for each category. The overall GISS Type is determined based on the individual finding representing the most severe injury (the right-most chosen column). Per the tool design, Types 1 and 2 constitute Class A findings (no tissue disruption) and Types 3, 4, and 5 are considered Class B (tissue disruption present). Fig. 2 shows an example of a completed GISS with its scoring.

## 5. Study objective

For such an instrument to have clinical, forensic, and/or evidence-based utility, it must be both valid and reliable. In order for an instrument used in evaluating genital injury after sexual intercourse to be considered reliable, it should yield similar results, regardless of the experience profile of the trained examiner.

One objective of this study is to determine inter- and intra-rater agreement of female external genital exam findings of examiners possessing a wide range of experience levels, utilizing the GISS. The combined objectives are to determine the reliability of the GISS as a data-collecting instrument and to extrapolate from that data, agreement levels regarding exam images and documentation, individually, and in combination. Additionally, we aim to determine which specific exam finding categories from the GISS are the most reliably agreed upon.

## 6. Sample characteristics

To examine the intra-rater reliability of the GISS, 49 female patients who had sexual assault examinations performed, had external genital images obtained, and for whom a GISS was completed, were selected. These 49 charts consisted of examinations conducted by six MLP forensic examiners (Group A), at least one year prior to initiation of the study. This was to minimize the chance that the raters would identify the cases they examined. The charts were chosen randomly from the emergency department's SA database with the criteria being that they were originally examined by one of the six Group A raters, that they had acceptable images taken at the time of the exam, and that a GISS was completed at the time of the exam process. The images were de-identified and given random identification numbers and were then randomly arranged.

## 7. Raters' characteristics

The initial six MLP examiners (Group A) who served as raters were stratified into three levels of experience. Two high level experience examiners each had greater than 20 years of experience performing sexual assault exams. Two moderate level experience examiners had between 11 and 15 years of experience. The two low level experience examiners had between 5 and 9 years of experience. All six of the raters had the same level of experience completing the GISS, as our SAC began utilizing it prospectively for all of its SA exams in 2007. A second group of six raters (Group B) with varying levels of experience performing exams and completing the GISS were utilized to score the combination of the images and documentation simultaneously for each of the 49 cases. The Group B cohort was formed to mitigate the potential unblinding of the study sample by Group A, who had previously reviewed and rated the study cases. In Group B, the highest experienced raters each had 8 years of experience scoring the GISS and greater than 13 years of experience performing exams. The moderate level raters had 7 and 8 years of GISS and exam experience, and the low experienced raters each had 3 years.

For the purpose of displaying study results, the 12 raters were assigned the following designations: high level raters for each cohort were named H1 and H2; moderate level experience raters, M1 and M2; and the low level experience raters, L1 and L2. See Fig. 3 for clarification of the raters' hierarchy and methodical flow.

## 8. Methods

The study sample was selected from cases performed at the