Proof of Concept for an International Long-Time Preservation ECG Format

Roberto Sassi[1], Luca Sparagino[1], Norman L Stockbridge[2], Juan M Guadiana[3], Fabio Badilini[4]

[1]Dipartimento di Informatica, Università degli Studi di Milano, Crema, Italy
[2]Division of Cardiovascular and Renal Products, FDA/CDER, Silver Spring, MD, USA
[3]Federal Aviation Administration, Las Cruces, NM, USA
[4]AMPS-LLC, New York, NY, USA

Abstract

We present a proof-of-concept of a medical data format for storing and retrieving 12-lead, 10-second standard electrocardiograms (ECG). The new format combines a digital ECG record into a standard graphical report (PDF).

Using the very recent PDF/A-3 standard (ISO 19005-3:2012), we designed a long-term preservation graphic report (such as a standard 3×4 or 6×2 ECG printout) in which the original ECG digital data (using the aECG ANSI/HL7 R1-2004 format) is embedded as a compressed attachment. The proposal also includes two ways of ensuring conformity between the ECG vector images and the digital data. First, a digital signature is added to guarantee that the document was not modified after creation. Second, specific metadata are inserted to allow a direct digital comparison between the vector ECG printout and the data.

To prove the feasibility of the approach, 93 aECGs, selected from different device manufacturers, resolutions and sample rates, were converted into the new format and then validated. Differences between digital data and vector graphics were always smaller than 0.02 μV, significantly below the typical A/D converter resolution.

1. Introduction

Most modern electrocardiographs digitally acquire standard 12-lead ECGs synchronously for 10 seconds. Depending on the manufacturer, the storage and further diffusion of the data may be more or less complex and may require the usage of an ECG management system.

Because of these logistical difficulties, mainly related to data portability, even in an era where a digital ECG is recognized and recommended as a suitable way to store and preserve records, and despite a number of standards being available [1], ECG paper reports remain a preferred and frequent choice of clinicians.

ECG systems by most manufacturers provide the possibility of exporting traditional 12-lead printouts in common formats, *e.g.*, ISO 32000-1:2008 Portable Document Format (PDF). These printouts permit some qualitative interpretation; however, most often only a portion of the data is included and, even worse, any analysis from the graphical picture is limited by the printing resolution of the system. On the contrary, raw digital ECG data would facilitate subsequent quantitative assessments, with a clear advantage in diagnosis.

More than 10 years ago, the need for fully reviewable ECGs led the US Food and Drug Administration (FDA) to require digital submission of ECG waveforms and annotations employed in the cardiac evaluations of new drugs. A specific industry-independent standard based on Health Level Seven (HL7) and structured in XML was designed by several partners leading to a format that encapsulates ECG waveforms and the associated annotations produced by an algorithm or by a human reader. This standard, commonly referred to as the *annotated ECG standard* (aECG), has been adopted by ANSI in May 2004 (ANSI/HL7 V3 ECG, R1-2004).

As of today, ECG records are thus either stored as paper or PDF reports, or in one of the digital standards available [1] (aECG is only one of the possible options). Thus they cannot be easily interchanged in the (frequent) case that clinical or research analyses may require it.

We introduce a proof of concept (PDF-ECG) that combines the benefits of a digital ECG and a standard graphical report. Our solution is based on the very recent PDF/A ISO 19005-3:2012 standard, which dictates an electronic document format for long-term preservation. The proof of concept designed is a PDF/A-3u (Unicode) envelope containing the ECG graphic (with no predefined layout), the ECG digital data in aECG format, and a digital signature. We also provide a means to ensure content match. PDF-ECG provides, within a unique structure, a way to satisfy both the needs for a simple graphic report, accessible without the installation of specific proprietary software, and for delivering all the acquired information for further processing with specialized software.

2. Methods

2.1. The HL7 aECG standard

The HL7 aECG standard was created in response to the FDA's digital electrocardiogram initiative introduced in 2001 [2]. The FDA required both ECG waveforms and annotations, generated for regulatory submission, to meet a specific format. At the time, no current standard for ECG waveforms met all the FDA's needs. As a result, the FDA, together with representatives from pharmaceutical industry, central core laboratories, and device manufactures worked together within Health Level Seven (HL7), a standards developing organization, to create a format to meet the FDA requirements.

The aECG standard was developed according to HL7's version 3 message development process (it passed final balloting in January 2004). Like other HL7 v3 formats, aECG is XML based, which means that the implementation of an aECG object must follow a specific XML *schema*, that defines in details the rules for a given file to be valid. These rules include the specific names of the XML mark-ups (tags, or elements), as well as which tags are mandatory and which are optional, according to an official Implementation Guide document [3].

While being strongly structured, and certainly with a significant amount of overhead (a standard 12-lead 10-second ECGs can easily reach the size of a few hundred kilobytes, compared to only a few kilobytes for other binary standards [1]), aECG has rapidly spread and was adopted by most manufacturers as well as by many research organizations. Part of the reason is certainly related to the driving force of the FDA, who has adopted the standard for the digital submission of ECG data from pharmaceutical studies, but also to the high practicality of the standard. In fact, HL7 wanted the standard to be as broad as possible, limiting the rules specified in the *schema* to include only the basic elements to store a generic digital ECG. Hence, also due to available software [4], aECG is today used for different purposes and not only for data submission to FDA (e.g., to share ECGs and patient's data with third parties, for research or clinical purposes).

In summary, even if it was developed to meet specific needs of US regulatory bodies, the aECG has been designed to be a broad multi-purpose ECG standard and resulted to be an ideal platform for data interchange which is today a reality.

2.2. The PDF/A-3u standard

PDF is a digital format, originally developed by Adobe Systems Inc., for representing printed documents. It was derived from a subset of the PostScript language, and while Adobe holds patents to PDF, it subsequently made the specifications publicly available. Also it cooperated with ISO to make an international standard (ISO 32000-1:2008) out of PDF 1.7.

As of today, PDF files are used to represent a large body of information around the world, and while this information must be preserved for a long time, multiple generations of technology might hinder accessibility. For these reasons, a second ISO standard (19005-1:2005) defines a file format based on PDF, known as PDF/A-1, which is meant for long term preservation of information, independently of the tools employed for creating or viewing the document. In practice, PDF/A-1 restricts the number of PDF components and features which might be used. Moreover, PDF/A-1 provides standard metadata to represent context, history and logical structure of the information within the document.

The PDF/A standard was extended over the years, and the latest recent revision (ISO 19005-3:2012), PDF/A-3, permits the embedding of files of any format (including XML, CSV, CAD, etc.).

2.3. The PDF-ECG proof-of-concept

PDF-ECG was conceived as an *hybrid archiving* format [5], where a source document in a less preservation-robust format is embedded within a PDF/A file. In specific, PDF/A-3u (Level U conformance) was selected to be Unicode compliant. Apart from classical word processing applications, similar approaches are being concurrently developed for other applicative contexts. For example, the specifications of the German ZUGFeRD format for electronic invoices were recently published [6]. In that case, the primary document component is a human-readable form of the invoice while the embedded XML file is intended for machine-processing.

The main content of a PDF-ECG file is a standard ECG report, such as (but not limited to) a 3×4 or 6×2 leads printout. No constraints are set on the graphic report, (except a few technical implementation details described below) and any manufacturer can shape it to its needs. The file can be opened with any standard PDF viewer, thus hiding the further complexity to users who might not need to access the digital data (*e.g.*, patients). The aECG file is embedded into the PDF/A file as a compressed object, reducing significantly its footprint. For hybrid archiving, PDF/A-3 mandates that an explicit association must be made between each *associated* embedded file and the PDF container (or one of its objects) by means of the *AFRelationship* key. Typically, such association is set to *Data* (meaning that the data can be employed to rebuild the PDF content) or *Alternative* (meaning that the two representations are equivalent). In PDF-ECG the AFRelationship was set to the latter. While the aECG data can be extracted using standard software by any user, its main purpose is to be employed by automated specialized algorithms or clinical tools used by clini-

Table 1. Results of the preliminary validation of the PDF-ECG format. The digital aECG XL7 files employed are subdivided by producer and characteristics.

aECG producer	f_s (Hz)	LSB (µV)	Files #	Validation (µV)
AMPS	180	6,25	5	0,01750
AMPS	250	4,88	3	0,01750
AMPS	500	1	5	0,01750
AMPS	500	2,5	4	0,01750
AMPS	500	4,878	5	0,01761
AMPS	500	4,88	7	0,01750
AMPS	500	5	2	0,01750
AMPS	1000	2,5	6	0,01750
GE	1000	2,5	12	0,01750
Mortara	1000	2,5	5	0,01750
Other	500	0,25	9	0,01750
Other	500	4,88	5	0,01750
Other	500	6,25	1	0,01750
Other	1000	0,5	3	0,01750
Other	1000	2,5	15	0,01750
Other	1000	3,75	5	0,01750
Other	1000	3,76	1	0,01750

cians to perform further analysis.

Unfortunately, there is no standard way to validate the relationship between the main content and any embedded files in PDF/A-3, as the National Digital Stewardship Alliance recently pointed out [5]. Therefore, even if the two representations were declared as *Alternative*, it would be possible, by mistake or deliberately, to produce a PDF-ECG document where the aECG data did not correspond to that displayed in the PDF file. To avoid this we designed a two-fold verification scheme. First, the institution or entity creating the PDF-ECG file has the option to sign the document digitally, to avoid subsequent modifications. Since the signature verification procedure is embedded in most viewers, its validation does not require additional software. Thus, the digital signature puts the correctness of the stated AFRelationship under the responsibility of the signer. However, there are situations in which this is not possible or inappropriate. Also, a digital signature does not protect from non-deliberated mistakes and technical problems. For this reason, we put a significant effort in constructing PDF-ECG such that it may be validated after its creation by a third party. Briefly:

- Report pages must contain a main layer. Its name must be "`LAYOUT_[hs]:[vs]`" where `hs` and `vs` are the horizontal (in mm/s) and vertical (in mm/mV) scales, respectively, *e.g.*, `LAYOUT_25:10` represents the classical 25 mm/s and 10 mm/mV ECG scales.
- For any aECG `<sequenceSet>`, a new layer is added (sibling of the main layout).
- For any HL7 signal, a further layer is created (child of the sequenceSet layer in which it is contained). The name of the layer describes its content according to the scheme:
  "`[CODE](counter*)_[s]:[i]:[e]:[offset]`"
  where `CODE` is the HL7 name of the signal, `s`, and `e` are the index of the first and last samples printed and `i` is the increment (in samples) employed. In fact, most often the ECG report only contains a portion of the information in the digital file. Also, `offset` is the vertical offset on the page, in postscript points. `counter*` (optional) is employed to enumerate signals printed more than once (*e.g.*, in the 3×4 format, `MDC_ECG_LEAD_II(2)_0:1:2600:325`)
- Every continuous ECG waveform, plotted on the report, must be contained in a single independent PDF stream contained in its layer.

At validation, after deflating the compressed PDF streams, the position of the points (which plot on the graphic report a ECG waveform) are extracted. The layer naming convention facilitates identifying each lead. Then, the vertical offset (`offset`) is removed and the positions of the points are rescaled to mV by using the scale extracted from the name of the main layer, *e.g.*

$$\mathrm{mVPoints_i} = \frac{25.4 \times (\mathrm{pdfPoints_i} - \mathrm{offset})}{72 \times 10}.$$

Finally the value of each $\mathrm{mVPoints_i}$ is compared with the corresponding sample in the embedded aECG file. PDF 1.7 uses single precision floats or 32-bit fixed point numbers, so the comparison in a genuine file will report a small difference $\ll 1$ µV, due to rounding errors (unless using the same identical numerical library, which is not practical). However, when the aECG embedded file and the PDF report differ, the difference is significantly larger than a few µV.

No validation is implemented at the moment for non-waveform data (*e.g.*, name of the patients, age, etc.), although it could be easily implemented in the future.

2.4. Data analysis

To assess the feasibility of the format, 93 digital ECGs, all formatted as aECG objects, were selected from a few ECG clinical studies which were previously submitted to the ECG Warehouse for FDA review. Table 1 lists the characteristics of the files. The selection was purposely heterogeneous in order to demonstrate flexibility and it included ECGs with different sampling rates and resolutions. Furthermore, aECG files were produced by different manufacturers.

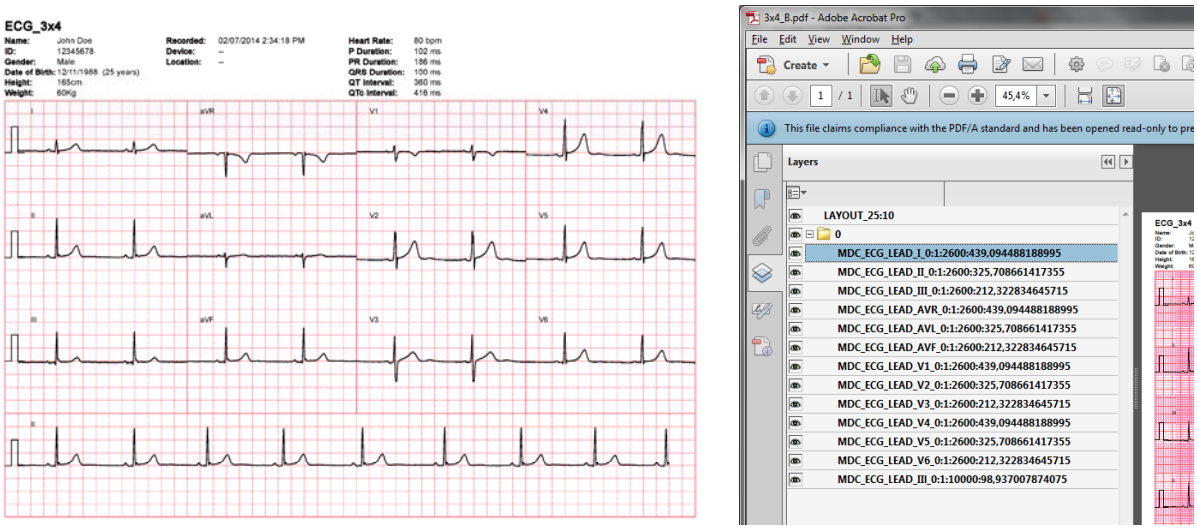

Figure 1. An example of 3×4 ECG report contained in a PDF-ECG file (left panel). To permit subsequent validations, the file contains a layer for each waveform, with a name prescribed by the format (right panel).

3. Results

ECG files were checked to be valid aECG objects with a commercial software package [7]. Then, using a C# application we developed, each digital file was rendered in a vector standard report (either 3×4 or 6×2). Figure 1(left) contains an example of report. A PDF/A-3u file encapsulating the digital data was prepared along the lines of the previous section. The manipulation of the PDF-ECG file was performed using a commercial library (PDFlib GmbH, Germany). The conformance to the PDF/A-3u format of the files was checked with the commercial Adobe Acrobat preflight tool.

A second C# ad-hoc application was employed to validate that the graphic report in the PDF-ECG and the HL7 file embedded corresponded up to machine precision. The last column of Table 1 reports, for each class of files, the corresponding maximum validation discrepancy, which was always smaller than 0.02 µV (28 times smaller than the finest resolution among the 93 files).

4. Conclusions

The new PDF-ECG document can be redistributed and opened with most available PDF readers, without specific software. Moreover, analysis software could be easily adapted to accept the file as input for subsequent analysis, which could be indeed overlaid on the printout itself. PDF-ECG long-term preservation characteristics make it suitable for long-term archival in hospital patients managements systems, core laboratory and research center databases. Finally, the PDF-ECG formats permits third-party validation, solving a problem which might prevent the adoption of this form of hybrid archiving.

The PDF-ECG format is still in a very preliminary stage, but we hope that manufacturers and institutional bodies might be interested in setting up a working group for its further formalization.

References

[1] Bond RR, Finlay DD, Nugent CD, Moore G. A review of ECG storage formats. Int J Med Inf 2011;80:681–697.

[2] Brown BD. The FDA's Digital ECG initiative and its impact on clinical trials. In: Morganroth J, Gussak I, editors. Cardiac Safety of Noncardiac Drugs, Humana Press 2005: 301–327.

[3] Brown BD, Badilini F. HL7 Version 3 implementation guide: regulated studies – annotated ECG, Release 1. Health Level Seven International 2005.

[4] Badilini F, Isola L. Freeware ECG viewer for the XML FDA format. Proceedings of the 2nd OpenECG Workshop, Berlin, Germany: 2004:. 31–34.

[5] NDSA Standards and Practices Working Group. The benefits and risks of the PDF/A-3 file format for archival institutions. National Digital Stewardship Alliance 2014.

[6] Forums, elektronische Rechnung Deutschland (FeRD). Das ZUGFeRD-Format. Spezifikation und Umsetzungsregeln zum branchenübergreifenden Kern-Rechnungsformat. Version 1.0 2014.

[7] Vaglio M, Isola L, Gates G, Badilini F. Use of ECG quality metrics in clinical trials. Comput Cardiol 2010;37:505–508.

Address for correspondence.

Roberto Sassi
Dipartimento di Informatica, Università degli Studi di Milano
via Bramante 65, 26013 Crema (CR) Italy
roberto.sassi@unimi.it