

## Improving the accuracy of MRI spleen and liver volume measurements: A phase III Gaucher disease clinical trial setting as a model

Luc Bracoud<sup>a,\*</sup>, Harris Ahmad<sup>b</sup>, Einat Brill-Almon<sup>c</sup>, Raul Chertkoff<sup>c</sup>

<sup>a</sup> BioClinica Inc., Lyon, France

<sup>b</sup> BioClinica Inc., Newtown, USA

<sup>c</sup> Protalix Biotherapeutics, Carmiel, Israel

### ARTICLE INFO

#### Article history:

Submitted 30 July 2010

Available online 17 November 2010

(Communicated by A. Zimran, M.D.,  
09 October 2010)

#### Keywords:

Gaucher disease

MRI

ERT

Spleen volume

Liver volume

### ABSTRACT

**Purpose:** To achieve minimal inter-observer variability in assessment of spleen and liver volume changes using a novel MRI reading method in the context of a phase III clinical trial of a new therapy for Gaucher disease.

**Materials and methods:** Abdominal MRI examinations at screening and after 6 and 9 months' exposure to a novel plant-cell-derived recombinant enzyme, taliglucerase alfa, were taken in 31 patients with Gaucher disease and at least 8-fold greater than expected splenomegaly. Transverse T2, T1, and in/out-of-phase, and coronal T1 sequences were performed using standardized settings across 11 sites globally. Spleen and liver volumes were semi-automatically delineated using an automatic segmentation algorithm followed by manual correction by experienced technologists using advanced editing tools. Data of all randomized patients were then submitted for efficacy evaluation to two independent experts blinded to time-point and treatment.

**Results:** Mean ( $\pm$ SD) percent variability over all time-points was  $0.30\% \pm 0.46\%$  for spleen and  $0.53\% \pm 0.69\%$  for liver using 178 spleen and liver volumes measured twice. Adjudication due to  $\geq 5\%$  variability between observers was not required.

**Conclusion:** The measurement method was found to be precise in monitoring spleen and liver volume changes over time, with a much lower variability than traditional manual methods, supporting the accuracy of the results. Given the observed minimal variability rates among multiple readers, a single read of each volume would be sufficient.

© 2010 Elsevier Inc. All rights reserved.

### Introduction

Gaucher disease, the most prevalent lysosomal storage disease [1–3], is caused by mutations in the glucocerebrosidase gene (GCD), leading to a reduced activity of the lysosomal enzyme glucocerebrosidase and an accumulation of the substrate glucocerebroside in the cells of the monocyte–macrophage system. This storage leads to hepatosplenomegaly and consequent hypersplenism as well as skeletal involvement, and less frequently, lung and neurological involvement [4].

Identification of GCD deficiency as the etiology of Gaucher disease stimulated the development of the first enzyme replacement therapy (ERT) as a therapeutic strategy for this and eventually other single gene lysosomal diseases. ERT provides sufficient exogenous active enzyme to degrade the accumulated substrate in the macrophages and ameliorate the disease-specific visceral features. ERT, which has been available for Gaucher disease for nearly two decades, has been shown to be safe and clinically efficacious within 2–5 years [5–8].

Clinical trials to introduce new ERTs or other therapeutic modalities in Gaucher disease have opted for multiple primary endpoints for proof of efficacy because of the need for stringent standardized criteria that are unambiguous. Since reduction in organomegaly is a clinically relevant feature of successful treatment that also translates into improvement in the hematological features, changes in spleen and liver volumes are common endpoints in clinical trials of Gaucher disease [9,10]. Nonetheless, in designing efficacy endpoints for clinical trials in Gaucher disease, there may be confounding variables when choosing improvement in anemia or thrombocytopenia (e.g., concurrent iron-deficiency or idiopathic thrombocytopenic purpura, respectively) or reduction in hepatomegaly (especially when it is very minimal). Hence, change in spleen volume is the most ideal parameter as a marker of efficacy, but absolute accuracy in measurement would have to be proven to be independent of observer bias in order to accommodate a multi-center trial.

This paper describes a novel method to achieve a very low degree of inter-observer variability in MRI spleen and liver volume assessments as demonstrated in a phase III clinical trial of a new ERT for Gaucher disease.

\* Corresponding author. BioClinica Inc., Bioparc, bât. Adénine, 60 av. Rockefeller, 69008 Lyon, France.

E-mail address: [luc.bracoud@bioclinica.com](mailto:luc.bracoud@bioclinica.com) (L. Bracoud).

## Materials and methods

### Study methodology

The study used as a prototype for the methodology a phase III randomized double-blind parallel-group dose-ranging clinical trial involving 11 international sites that received IRB approval and recruited 31 treatment-naïve patients with Gaucher disease who consented to receive a plant-cell-derived recombinant ERT (taliglucerase alfa; Protalix, Carmiel, Israel). The single primary endpoint was reduction in spleen volume after 9 months of treatment in two-dose groups; liver reduction was a secondary endpoint [11,12]. The major entry criterion beyond age and disease status was splenomegaly greater than 8 times the volume expected by body weight [13].

### MRI data

Up to 3 MRI examinations (screening and after 6 and 9 months' exposure to ERT) were performed in each patient using standardized MRI acquisition settings in all 11 global sites:

- T2-weighted Transverse (Turbo Spin-Echo, Respiratory Triggered when available)
- T1-weighted Transverse (3D preferred, e.g. VIBE, LAVA or 3DTFE)
- T1-weighted Coronal (3D preferred, e.g. VIBE, LAVA or 3DTFE)
- Dual Gradient-Echo In/Out-of-phase Transverse.

Data were acquired on Philips Intera (3), Philips Achieva (1), GE Signa Excite (2), GE Signa HDx (1), Siemens Symphony (2), or Siemens Avanto (2) 1.5 T scanners.

All MRI data were sent for central processing to BioClinica Inc. (Lyon, France). Detailed quality control (QC) was performed to check protocol compliance and overall image quality. Repeat scans were requested in cases of QC failure: the most frequent reasons being incomplete organ coverage, strong in-plane motion, moderate through-plane motion, and inconsistent MRI settings across examinations. Of the 11 examinations that were required to be partly redone, only one failed to be successfully repeated.

### Volumetric assessment methodology

Spleen volume was delineated on each T2-weighted transverse sequence, while liver volume was delineated on each T1-weighted transverse sequence. The T1-weighted coronal and dual gradient-echo in/out-of-phase transverse sequences provided additional anatomical information. The transverse orientation was chosen for volumetric measurements since it is less prone to partial volume effects and breathing artifacts [14].

A fully automated, 3-D image segmentation algorithm based on Bayesian classification of MRI signal intensities and Markov Random Field models were used to classify voxels into tissue classes (Fig. 1, column 2).

The segmentation output was provided to experienced MRI technologists for manual correction using advanced editing tools. These technologists were chosen for their broad experience in performing volumetric quantification, including spleen and liver volumes, on MRI and CT data from various indications.

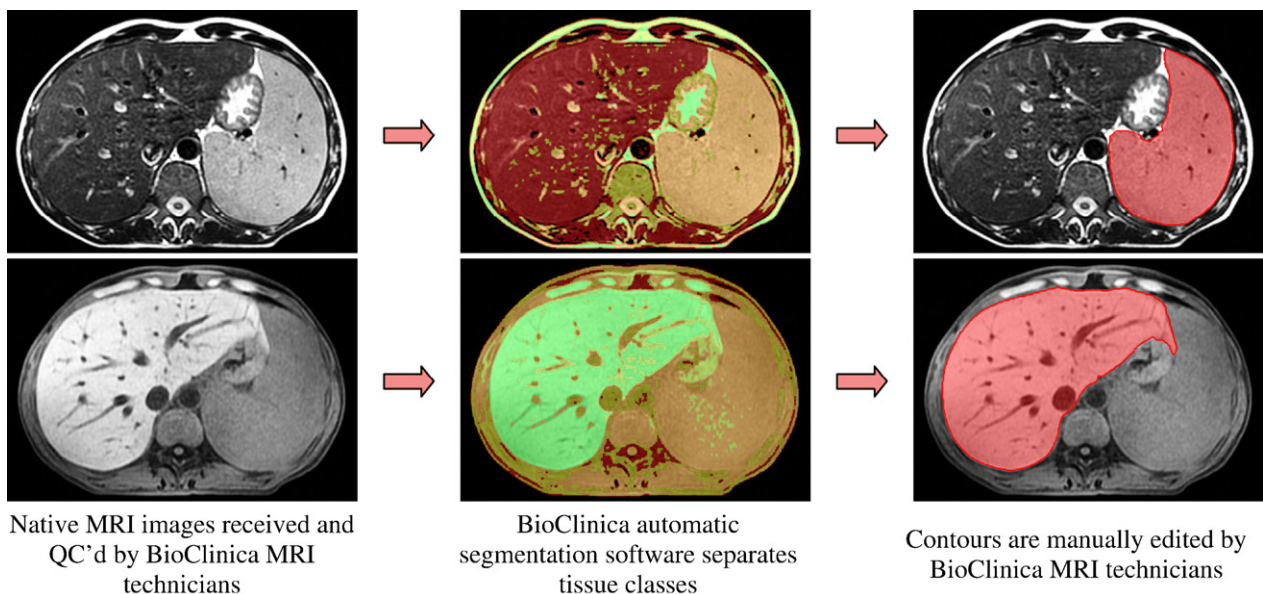
The technologists evaluated the spleen and liver tissue visually to remove false positives, differentiate from surrounding structures with similar signal intensity, and append incorrectly assigned segments (Fig. 1, column 3). The robust, validated segmentation software minimized further adjustments. When manual editing was required, a few reference points could be placed along the border of the organ and the system interpolated a curve to use as a boundary to add or remove the incorrectly detected areas. Cardinal spline-interpolated curves [15] were found to closely follow the actual organ contours and facilitated smoother contouring.

The resulting contours of each organ at each time-point were reviewed by independent central imaging experts blinded to the treatment group, patient ID, and time sequence, to obtain the final organ contours (Figs. 2 and 3).

The technologists as well as the blinded readers were previously trained on the editing tools and image review system and their intra- and inter-observer variability was assessed on non-study subjects to ensure reliability.

### Central image review strategy

The data were displayed in a dedicated image review system (Fig. 4) using a proprietary user-friendly environment that allowed



**Fig. 1.** Semi-automatic contour detection. Spleen contours are detected on T2-weighted data (first row), while liver contours are detected on T1-weighted data (second row). The first column shows the native images. The second column displays the output of the automatic segmentation as an overlay to the image data. The third column shows the final organ contours after manual validation.

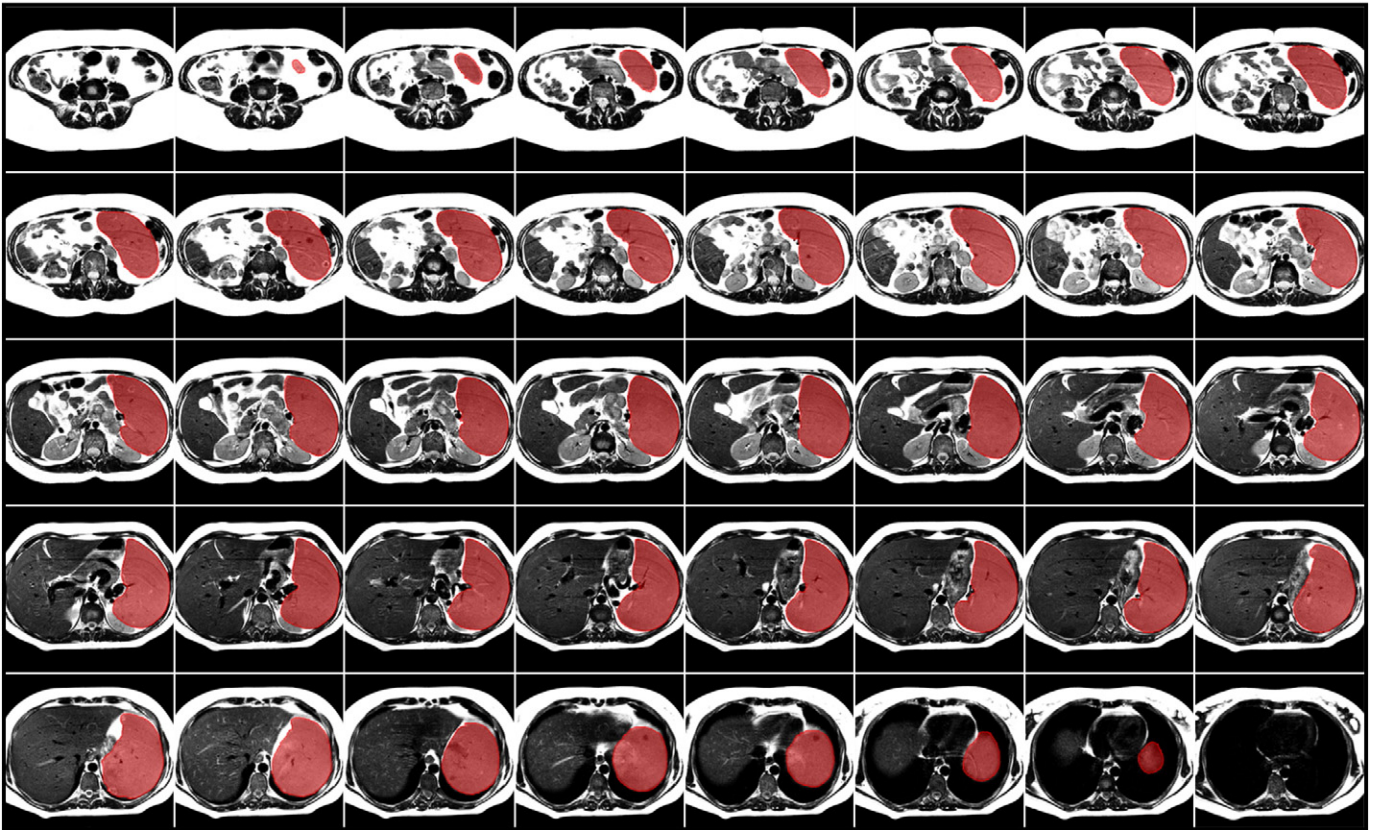


Fig. 2. Example of spleen contours after automatic segmentation and manual validation. Validated spleen contours are overlaid on the native T2-weighted transverse images on each relevant slice.

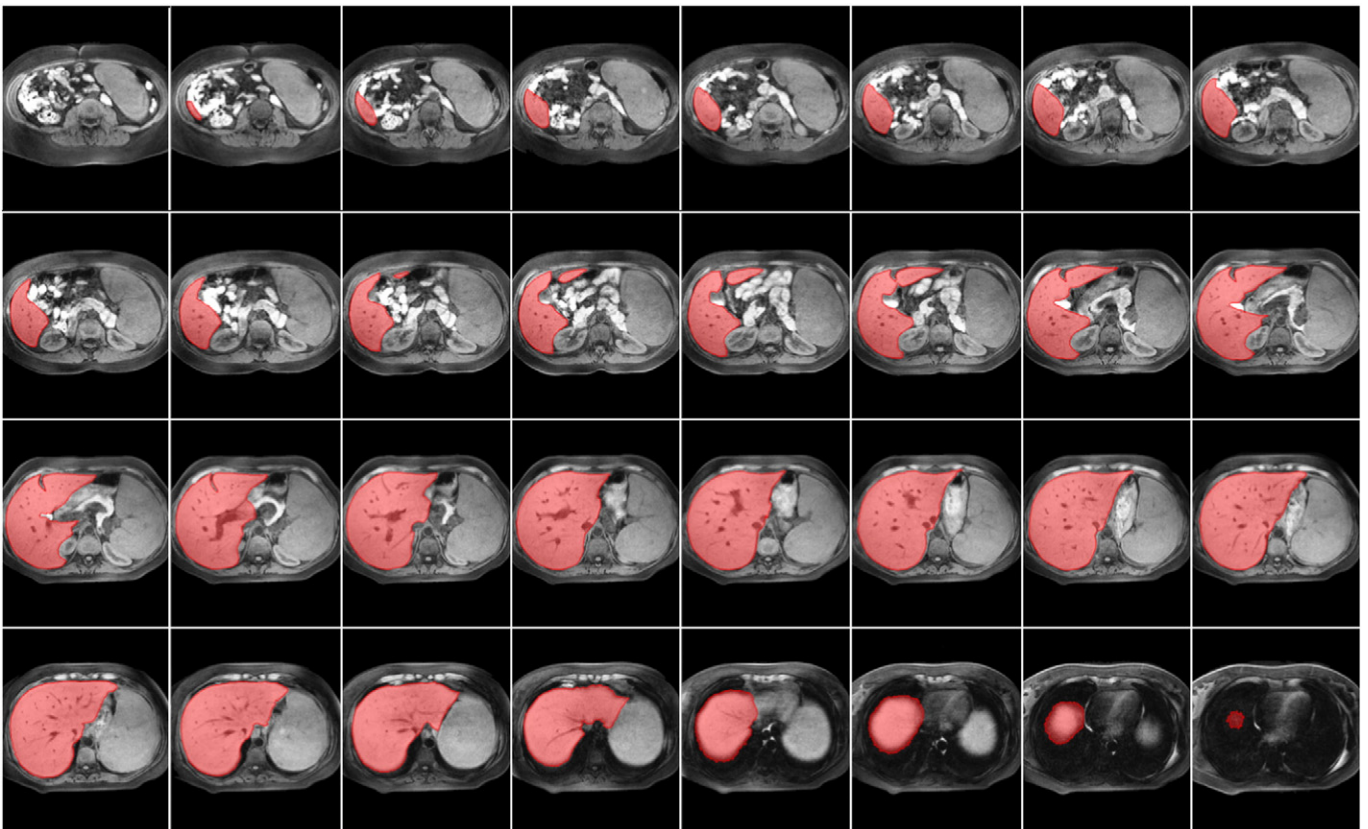
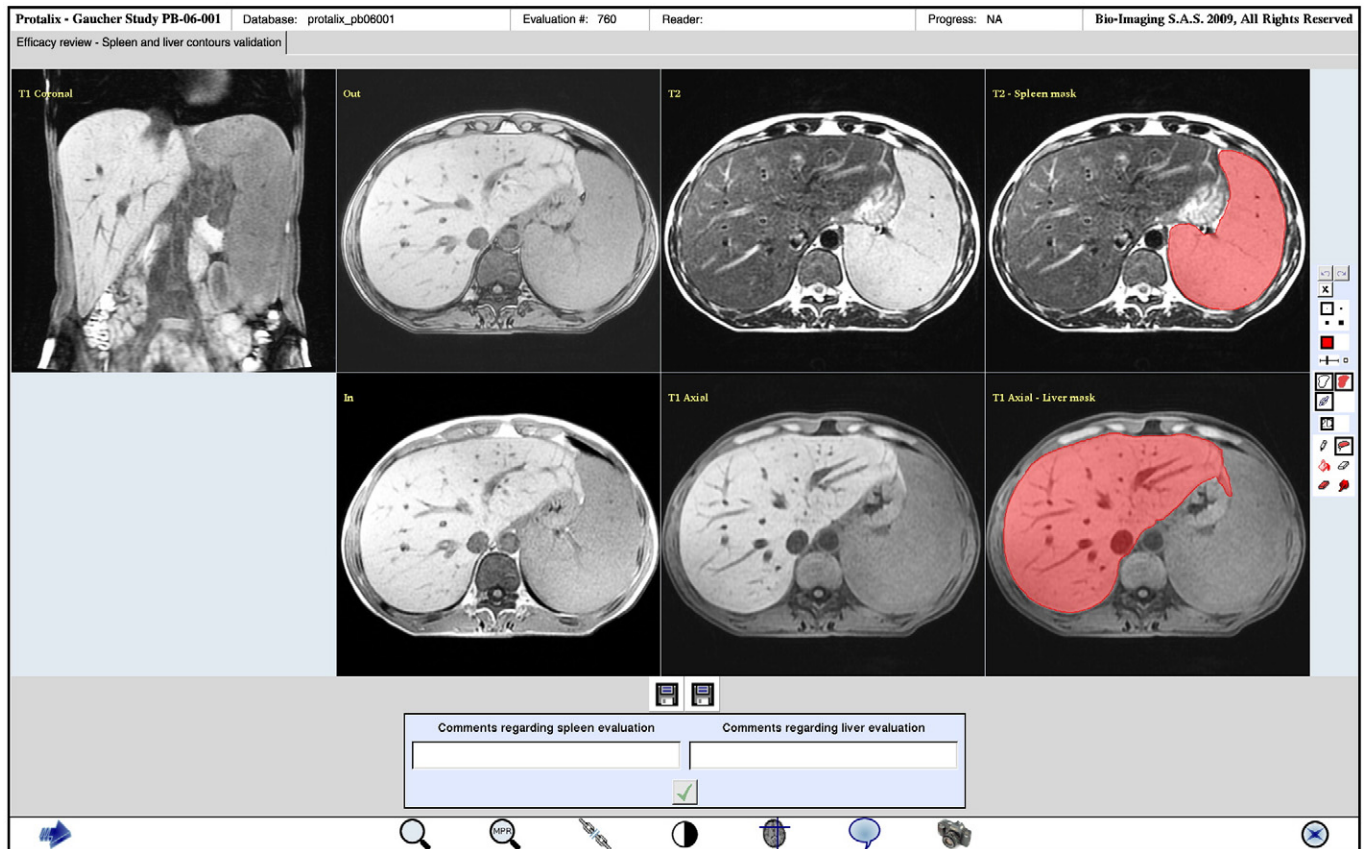


Fig. 3. Example of liver contours after automatic segmentation and manual validation. Validated liver contours are overlaid on the native T1-weighted transverse images on each relevant slice.



**Fig. 4.** Screenshot of the Image Review Software used for the validation of spleen and liver contours. All MRI sequences available for a given examination are juxtaposed and the organ contours are overlaid on the sequences used during the pre-detection for validation.

review of all available MRI sequences as well as interactive realignment and multi-planar reconstruction. Once organ contours were validated and saved into the central database, organ volumes (in milliliters) could be obtained as the sum of the areas from each slice multiplied by slice thickness using Simpson's rule.

All evaluable data were submitted to two readers for efficacy evaluation. In case of significant discrepancies between the readers ( $\geq 5\%$ ), an adjudicator was to be involved. The final value for efficacy evaluation was either the mean volume of the two readers or the adjudicated volume.

#### Variability assessment

Variability ( $V$ ) was measured as the relative percentage difference between the two measurements ( $Vol1$  and  $Vol2$ ) for each of the organs reviewed twice, using the below formula:

$$V(\%) = \frac{|Vol2 - Vol1|}{\text{Mean}(Vol1, Vol2)} \times 100 \quad (1)$$

## Results

#### Variability results

The variability in 178 volumes (89 each of spleen and liver) reviewed twice is presented in Fig. 5.

The mean inter-observer variability per time-point is summarized in Table 1. Overall, mean variability in spleen measurements was 0.30% compared to 0.53% for liver measurements.

Variability being less than 5% for all cases, no adjudication was required.

The outlier which can be seen at screening on both spleen and liver measurements corresponds to an examination of poor quality which could not be repeated.

#### Causes of variability

The most frequent (albeit minimal) sources of variability between readers were:

- Partial volume effect
- Separation with surrounding structures: kidney and stomach for spleen; kidney, pancreas, gall bladder, and heart for liver
- Delineation of upper slices
- Inclusion/exclusion of extra-parenchymatous vessels.

Maximizing image quality was important in order to facilitate the organ delineation.

## Discussion

#### Factors influencing accuracy

Given the large variety of scanners used across sites, which is inevitable in the context of an international multi-center trial, the accuracy of a cross-sectional analysis may be decreased, unless one is able to monitor the variability induced by the systems themselves and make sure it is minimal. This could be addressed by scanning abdominal phantoms multiple times at different sites. The influence of field strength would also be a subject of interest, even though here all scans were performed on 1.5 T scanners.

But what matters most for such a clinical trial is the longitudinal follow-up of subjects. It is therefore important to make sure that each subject is followed on the same MRI system during the course of the

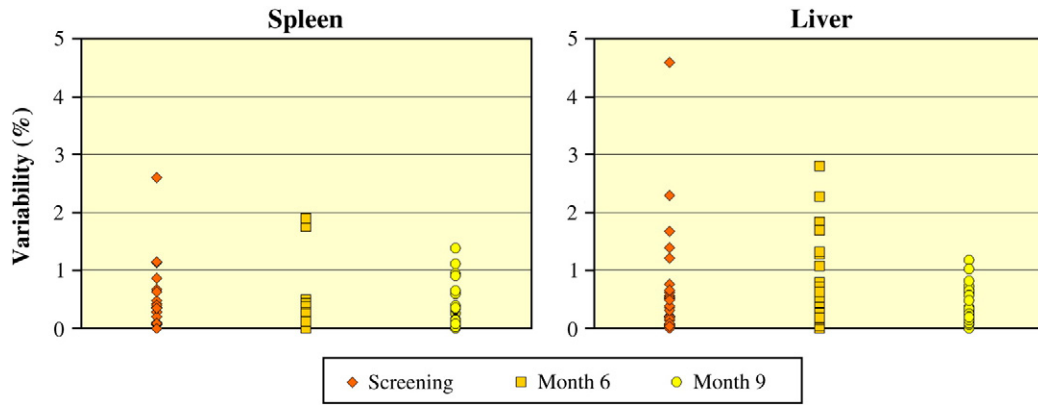


Fig. 5. Detailed variability results. Mean inter-observer variability for spleen and liver volume readings at each time-point.

study, using the same acquisition parameters, and minimize the changes made to that system. Nevertheless, system maintenance, software or hardware upgrades, or even system replacements cannot always be avoided if a study lasts for many years. Again, scanning an abdominal phantom before and after the system modification could help assess the effects of such changes.

Another element which could influence the evaluation is fasting. Although the impact of fasting on spleen and liver volumes has never been precisely quantified on humans, it could be recommended for future studies to ask the subjects to fast for a specific time period prior to the MRI examination, to make sure that the data are more homogeneous.

#### Advantages with respect to other existing methods

Spleen and liver volumes can be measured not only using MRI but also with computed tomography (CT) and ultrasonography (US). Previous studies have shown that each of these modalities provide with a reliable way of measuring such volumes [16,17]. Compared to CT, for an almost similar cost, MRI provides with a better way of visualizing soft tissues in detail without exposing the subjects to radiation. US remains the method of choice for assessing spleen and liver volumes routinely, given its wide access and cheaper cost. Nevertheless, such method relies on the measure of 3 maximal dimensions (antero-posterior, longitudinal, and transverse) using data of poor spatial resolution, in order to assess the volume of organs of complex shape with strong variations between subjects. For a study whose one of the eligibility criteria and whose primary efficacy endpoint is spleen volume, it therefore appears extremely important to make use of the best modality available in order to perform these assessments, and thus select MRI. For long term follow-up, subjects may then be switched to US.

The novelty in this work does not reside in the concept *per se* of monitoring spleen and liver volumes during the course of a study using MRI but in the systematic, centralized, and thorough review that was made. In the context of a multi-center trial, it makes sense that such analyses are externalized to a central laboratory with dedicated expertise, tools, and staff.

Table 1

Mean  $\pm$  SD % inter-observer variability for spleen and liver volume readings at each time-point.

	Spleen	Liver
Screening, %	0.35 $\pm$ 0.54	0.59 $\pm$ 0.91
Month 6, %	0.26 $\pm$ 0.44	0.66 $\pm$ 0.69
Month 9, %	0.28 $\pm$ 0.39	0.35 $\pm$ 0.30
Overall, %	0.30 $\pm$ 0.46	0.53 $\pm$ 0.69

Theoretically, spleen and liver volume measurements could be performed on any transverse sequence of the protocol. It was observed that tissue contrast gave good pre-detection results for spleen on T2-weighted images and for liver on T1-weighted images, thus the decision to perform the measurements on these sequences and to stick to it for the entire cohort.

The low inter-observer variability is primarily achieved because of the image segmentation algorithm which leads to minimal manual intervention on the pre-detected contours from the automatic segmentation. Factors that supported this minimal variability were ongoing QC of images received during the study, the validated rigor of the automatic segmentation software, and finally, experienced MRI technicians and imaging expert readers trained in proper delineation of organ contours prior to the advent of the study.

Inter-observer variability was found to be higher for liver than spleen, which can probably be explained here by the fact that the images used for liver delineation had a thinner slice thickness than for the spleen delineation, resulting in potentially more manual editing.

#### Conclusion

A semi-automatic organ volume measurement methodology utilizing an automatic segmentation software method was found to be precise in monitoring spleen and liver volume changes over time, with a much lower variability than traditional fully manual methods. Given the observed minimal variability rates among multiple central imaging expert readers, a single read of each case would be sufficient to accurately measure spleen and liver volumes, and detect response to treatment on follow-up examinations.

The same methodology is currently in use for evaluating the volume of tissue likely not to respond to ERT (infarcted tissue, fibrosis, Gaucher nodules) in this cohort of patients. Correlating these results to the efficacy results may be of great interest in order to better understand why some patients better respond to treatment than others.

#### Disclosures

As a contract research organization (CRO), BioClinica was selected by Protalix for managing the MRI component of their phase III study. The services performed were invoiced as stated in the CRO contract. BioClinica staff and the central imaging expert readers remained blinded to treatment groups and declared no conflicts of interest in this work. Besides, Einat Brill and Raul Chertkoff are full time Protalix employees.

## References

- [1] E. Beutler, G.A. Grabowski, Glucosylceramide lipidoses: Gaucher disease, in: C.R. Scriver, A.L. Beaudet, W.S. Sly, D. Valle (Eds.), *The Metabolic Basis of Inherited Disease*, 7th ed., McGraw-Hill, New York, 1995, pp. 2641–2670.
- [2] R.E. Lee, The pathology of Gaucher disease, *Prog. Clin. Biol. Res.* 95 (1982) 177–217.
- [3] G.A. Grabowski, Gaucher disease: enzymology, genetics, and treatment, *Adv. Hum. Genet.* 21 (1993) 377–441.
- [4] G.A. Grabowski, R.J. Hopkin, Enzyme therapy for lysosomal storage disease: principles, practice, and prospects, *Annu. Rev. Genomics Hum. Genet.* 4 (2003) 403–436.
- [5] N.J. Weinreb, J. Charrow, H.C. Andersson, et al., Effectiveness of enzyme replacement therapy in 1028 patients with type 1 Gaucher disease after 2 to 5 years of treatment: a report from the Gaucher registry, *Am. J. Med.* 113 (2002) 112–119.
- [6] A. Zimran, G. Altarescu, et al., Phase 1/2 and extension study of velaglucerase alfa replacement therapy in adults with type 1 Gaucher disease: 48-month experience, *Blood* 115 (2010) 4651–4656.
- [7] M. de Fost, C.E. Holla, et al., Superior effects of high-dose enzyme replacement therapy in type 1 Gaucher disease on bone marrow involvement and chitotriosidase levels: a 2-center retrospective analysis, *Blood* 108 (2006) 830–835.
- [8] L.W. Poll, J.A. Koch, et al., Correlation of bone marrow response with hematological, biochemical, and visceral responses to enzyme replacement therapy of nonneuropathic (type 1) Gaucher disease in 30 adult patients, *Blood Cells Mol. Dis.* 28 (2002) 209–220.
- [9] N.J. Weinreb, M.C. Aggio, et al., International Collaborative Gaucher Group (ICGG), Gaucher disease type 1: revised recommendations on evaluations and monitoring for adult patients, *Semin. Hematol.* 41 (Suppl. 5) (2004) 15–22.
- [10] D. Elstein, S. vom Dahl, Gaucher disease, in: A. Schattner, H. Knobler (Eds.), *Metabolic Aspects of Chronic Liver Disease*, Nova Medical Publishers Inc., 2007, pp. 225–243.
- [11] A. Zimran, M. Petakov, et al., Novel Enzyme Replacement Therapy for Gaucher Disease: Phase III Pivotal Clinical Trial with Plant Cell Expressed Recombinant Glucocerebrosidase (prGCD)—taliglucerase alfa, 9th International EWGGD Meeting 2010, Cologne Germany, 2010.
- [12] G.M. Pastores, N.J. Weinreb, H. Aerts, et al., Therapeutic goals in the treatment of Gaucher disease, *Semin. Hematol.* 41 (suppl. 5) (2004) 4–14.
- [13] J. Ludwig, *Current Methods of Autopsy Practice*, 2nd ed. W.B. Saunders, Philadelphia, 1979 p. 676.
- [14] E. Beutler, A. Demina, et al., The clinical course of treated and untreated Gaucher disease. A study of 45 patients, *Blood Cells Mol. Dis.* 21 (1995) 86–108.
- [15] E. Catmull, R. Rom, A class of local interpolating splines, in: R.E. Barnhill, R.F. Riesenfeld (Eds.), *Computer Aided Geometric Design*, Academic Press, New York, 1974, pp. 317–326.
- [16] D. Glenn, D. Thurston, et al., Comparison of magnetic resonance imaging and ultrasound in evaluating liver size in Gaucher patients, *Acta Haematol.* 92 (1994) 187.
- [17] D. Elstein, I. Hadas-Halpern, et al., Accuracy of ultrasonography in assessing spleen and liver size in patients with Gaucher disease: comparison to computed tomographic measurements, *J. Ultrasound Med.* 16 (1997) 209–211.