Molecular Genetics and Metabolism Reports 27 (2021) 100725

Contents lists available at ScienceDirect

Molecular Genetics and Metabolism Reports

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

The Gaucher earlier diagnosis consensus point-scoring system (GED-C PSS): Evaluation of a prototype in Finnish Gaucher disease patients and feasibility of screening retrospective electronic health record data for the recognition of potential undiagnosed patients in Finland

Markku J. Savolainen [a], Antti Karlsson [b], Samppa Rohkimainen [c], Iiro Toppila [d], Mariann I. Lassenius [d], Carlos Vaca Falconi [e], Kristiina Uusi-Rauva [d], Kaisa Elomaa [f,*]

[a] Oulu University Hospital, PO Box 10, 90029 OYS, Oulu, Finland
[b] Auria Biobank, Turku University Hospital, University of Turku, PO Box 52, 20521 Turku, Finland
[c] Biobank Borealis of Northern Finland, PO Box 50, 90029 OYS, Oulu, Finland
[d] Medaffcon Oy, Tietäjäntie 2, 02130 Espoo, Finland
[e] Takeda Pharma AB, Vasagatan 7, 11120 Stockholm, Sweden
[f] Takeda Oy, Ilmalantori 1, 00101 Helsinki, Finland

ARTICLE INFO

Keywords:
Biobank study
Electronic health record
Gaucher disease
Gaucher earlier diagnosis consensus point-scoring system
GBA
Lyso-Gb1

ABSTRACT

Background: Gaucher disease (GD) is a rare inherited multiorgan disorder, yet a diagnosis can be significantly delayed due to a broad spectrum of symptoms and lack of disease awareness. Recently, the prototype of a GD point-scoring system (PSS) was established by the Gaucher Earlier Diagnosis Consensus (GED-C) initiative, and more recently, validated in Gaucher patients in UK. In our study, the original GED-C PSS was tested in Finnish GD patients. Furthermore, the feasibility of point scoring large electronic health record (EHR) data set by data mining to identify potential undiagnosed GD cases was evaluated.

Methods: This biobank study was conducted in collaboration with two Finnish biobanks. Five previously diagnosed Finnish GD patients and ~ 170,000 adult biobank subjects were included in the study. The original PSS was locally adjusted due to data availability issues and applied to the Finnish EHR data representing special health care recordings.

Results: All GD patients had high levels of the biomarker lyso-Gb1 and deleterious *GBA* mutations. One patient was a compound heterozygote with a novel variant, potentially pathogenic mutation. Finnish EHR data allowed the retrospective assessment of 27–30 of the 32 original GED-C signs/co-variables. Total point scores of GD patients were high but variable, 6–18.5 points per patient (based on the available data on 28–29 signs/co-variables per patient). All GD patients had been recorded with anaemia while only three patients had a record of splenomegaly. 0.72% of biobank subjects were assigned at least 6 points but none of these potential "GD suspects" had a point score as high as 18.5. Splenomegaly had been recorded for 0.25% of biobank subjects and was associated with variable point score distribution and co-occurring ICD-10 diagnoses.

Discussion: This study provides an indicative GED-C PSS score range for confirmed GD patients, also representing potential mild cases, and demonstrates the feasibility of scoring Finnish EHR data by data mining in order to screen for undiagnosed GD patients. Further prioritisation of the "GD suspects" with more developed algorithms and data-mining approaches is needed.

Funding: This study was funded by Shire (now part of Takeda).

Abbreviations: DBS, dried blood spot; EHR, Electronic health record; GBA1/GBA, β-glucocerebrosidase gene; GD, Gaucher disease; GlcCer, β-glucosylceramide; GlcCerase, β-glucosylceramidase; GlcSph/Lyso-Gb1, β-glucosylsphingosine; GED-C, The Gaucher Earlier Diagnosis Consensus; HDSF, Hospital District of Southwest Finland; NOHD, Northern Ostrobothnia Hospital District; PSS, Point-scoring system.

* Corresponding author at: Nordic Innovation CoE Lead, Takeda, PO Box 1406 (Ilmalantori 1), 00101 Helsinki, Finland.
E-mail addresses: markku.savolainen@oulu.fi (M.J. Savolainen), aspkar@utu.fi (A. Karlsson), samppa.rohkimainen@ppshp.fi (S. Rohkimainen), iiro.toppila@medaffcon.fi (I. Toppila), mariann.lassenius@medaffcon.fi (M.I. Lassenius), carlos.vaca-falconi@takeda.com (C.V. Falconi), kristiina.uusi-rauva@medaffcon.fi (K. Uusi-Rauva), kaisa.elomaa@takeda.com (K. Elomaa).

https://doi.org/10.1016/j.ymgmr.2021.100725
Received 21 December 2020; Received in revised form 25 January 2021; Accepted 25 January 2021
Available online 9 February 2021

1. Introduction

Gaucher disease (GD) is an autosomal recessive lysosomal storage disorder caused by the deficiency of the lysosomal enzyme β-glucosylceramidase (GlcCerase; EC3.2.1.45; also referred to as acid β-glucosidase and β-glucocerebrosidase), required for the degradation of the cell membrane sphingolipid β-glucosylceramide (GlcCer) [1–3]. A consequent lysosomal accumulation of GlcCer in tissue-resident long-lived macrophages ("Gaucher cells"), in liver, spleen, and bone marrow in particular, is considered as a hallmark of GD. In addition to GlcCer, a downstream metabolic product of GlcCer, β-glucosylsphingosine (GlcSph or lyso-Gb1), as well as other molecular factors may accumulate and have a role in the pathophysiology of GD [3–6].

More than 300 mutations distributed throughout the GlcCerase-encoding *GBA* gene (glucosidase, beta, acid; MIM# 606463; also known as *GBA1*) have been reported with the majority being missense mutations, including the most prevalent mutations c.1226A > G (p.Asn409Ser, also known as N370S) and c.1448 T > C (p.Leu483Pro, also known as L444P) [7–9]. Pathogenic *GBA* mutations affect the stability and activity of the GlcCerase, but many have remained functionally unconfirmed [10]. The genotype-phenotype correlation has largely remained elusive due to complexity arising from variable phenotypic consequences of identical genotypes and conversely, genetic heterogeneity among clinically similar patients. Effects of a mutation(s) in a second allele, allelic complexity, and potential genetic and environmental modifiers may each play a role [8–10].

GD is a multisystem disorder with a wide phenotypic spectrum. Typical manifestations at diagnosis include thrombocytopenia, splenomegaly, hepatomegaly, bone or joint pain, and anaemia [11]. GD has traditionally been categorised into three main subtypes [3,12], although intra-type variation along with the recognition of new genetic variants and clinical manifestations has challenged the traditional categorisation [8,13–16]. Type 1 GD (MIM# 230800) is the most common subtype representing >90% of the GD cases, and varies from asymptomatic to early-onset disease, does not include neurological symptoms, and rarely is life-threatening [3,12]. The rarer GD subtypes, type 2 (MIM#230900) and type 3 (MIM#231000), represent more severe, early-onset forms of the disease, typically involving moderate to severe neurological symptoms. Type 2 patients usually die within the first two years of life. Type 3 has similar but more chronic and slower progressing disease course than type 2 GD [3,12]. In addition, GD patients have an elevated risk of Parkinson's disease and malignancies [17–19].

Although GD is rare, it is the most common lysosomal storage disorder. The prevalence of type 1 GD is estimated to be 1:30,000–40,000 in the general population and ~ 1:1,000 among Ashkenazi Jewish, while type 2 and type 3 occur with a frequency of approximately 1:100,000 [20]. Studies from Australia and France have reported on population-specific prevalence of 1:57,000 and 1:136,000, respectively [21,22]. Currently, there are approximately only 17 GD patients in Finland (personal communication, Prof. Markku Savolainen, Oulu University Hospital, Oulu, Finland), responding to a prevalence of ~1:325,000. This may suggest that a number of Finnish GD patients have remained unidentified.

GD can be diagnosed by measuring GlcCerase activity from blood or skin fibroblasts [3]. Genetic tests can be performed to identify the specific *GBA* mutations in question as well as possible carriers among the family members of affected individuals [3]. Patients are also tested for haematological abnormalities (anaemia, thrombocytopenia, or signs of liver dysfunction), bone abnormalities (X-ray), and hepato- and/or splenomegaly (MRI or CT scans) [3]. Nevertheless, the diagnosis of GD is often significantly delayed due to lack of disease awareness or misdiagnoses owing to the varying spectrum and level of symptoms that overlap with several other conditions [11]. According to Mehta et al. [11], more than half of the patients receive their diagnosis after 18 years of age, and as many as one of six patients receive diagnosis after seven years or more since first consulting a doctor.

The Gaucher Earlier Diagnosis Consensus (GED-C) initiative has recently reached consensus regarding signs and co-variables classified as major or minor early indicators of type 1 and type 3 GD [23]. One of the major indicators of GD was splenomegaly which reached 100% consensus among the GED-C panellists [23]. Together, the signs and co-variables suggested by the GED-C panel can be used as the prototype of a GD point-scoring system (GED-C PSS) to facilitate the guidance for early diagnostic testing across clinical disciplines. However, the prototype needs to be evaluated and its reproducibility tested in a real-world setting. Recently, the PSS adapted from the original GED-PSS was tested in 25 GD patients in UK [24].

Finnish GD patients are being investigated for their symptoms at university and central hospitals with electronic health record (EHR) data accessible for clinical and research purposes. EHR data can be linked to respective biological samples via national biobanks. In this biobank study, the prototype of the GED-C PSS was tested in five Finnish GD patients. Furthermore, the feasibility of using Finnish EHR data for point scoring signs/co-variables indicative of GD was assessed. The overall aim of this and a future follow-up study is to determine if point scoring of EHR data by data-mining tools followed by analysis of available biobank samples for diagnostics is a more effective means to identify potential GD patients than the current clinical workup.

2. Materials and methods

2.1. Ethics and study design

This retrospective-prospective biobank study governed by the Finnish Biobank Law 688/2012 was conducted in a collaboration with the Biobank Borealis of Northern Finland (Oulu, Finland) and the Auria Biobank (Turku, Finland).

Five previously diagnosed type 1 GD patients at Oulu University Hospital, Northern Ostrobothnia Hospital District (NOHD; or PPSHP in Finnish), Finland, and ~ 170,000 adult biobank sample donors, previously treated at the Hospital District of Southwest Finland (HDSF; or VSSHP in Finnish), whose EHR data obtained as part of previous hospital visits were accessible via Biobank Borealis and Auria Biobank, respectively, were included in the study. At Auria Biobank, only consents from adults were obtained at the time of the study. GD patients at the NOHD were diagnosed throughout 1994–2018 and are recorded according to the International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10) code E75.2 in the EHR.

GD patients provided a signed informed consent for participation in this study. Tentative study approval from Borealis was obtained on August 31, 2018 and the final approval from the ethics committee on October 8, 2018. Physician contacted the patients to inform about the study, after which patient information sheets and consent forms for biobank research (general and study specific consents) were sent to individuals by biobank personnel. Six out of seven contacted patients were reached, and five of the six reached were willing to participate the study (patients GD_1, GD_2, GD_5, GD_6, and GD_7).

Regarding the permission to use the data accessible via Auria Biobank, a study protocol and a data request were submitted to Auria for approval by the scientific steering committee. The committee approved the request on February 6, 2017.

2.2. GBA sequencing and lyso-Gb1 measurement

GBA sequencing and lyso-Gb1 measurements were performed at Centogene AG (Rostock, Germany) as previously described [6,25,26]. The sequencing was performed on dried blood spots (DBSs) and lyso-Gb1 levels were determined from DBS and plasma samples.

Samples were collected by the Biobank Borealis research nurses with four of five patients being on the enzyme replacement therapy (ERT) at the time of sample donation. DBSs were sent at room temperature while plasma was separated fresh and sent on dry ice to Centogene.

DNA sequence data was not submitted to GenBank. Study permission was based on informed consent, privacy policy, and material transfer agreement which do not allow the transfer of the data to other registries or outside of EU and ETA. Furthermore, only GBA aberrations were allowed to be collected and published.

2.3. Point scoring

The overall point score and distribution per sign/co-variable of GD was inspected in a longitudinal, retrospective EHR data using the indicators of the original GED-C PSS [23] with the local adjustments to match the local data format, Finnish language, and data availability (Table A.1). The original GED-C PSS was proposed by the experts of the GED-C initiative independently and outside of this study.

In our study, accessed data included special health care recordings on clinical diagnoses, laboratory measurements, operations, imaging results, pathology diagnoses, basic demographics, and medical charts' texts. The two assessed cohorts were independently point scored at the respective participating biobanks. Point scoring of five GD patients of the NOHD and the biobank population of the HDSF was carried out at the Biobank Borealis and Auria Biobank, respectively. Point scoring of GD patients was carried out to represent the period before start of the treatment of GD. For signs/co-variables that were based on laboratory tests, the status and score were determined based on the most prevalent status. In the GED-C point scoring of the biobank population, the data and the most prevalent status of laboratory results of adult subjects, available by April 2017, was utilised.

In both point-scoring assessments, data on Jewish ancestry was unavailable. Information regarding blood relative who died of foetal hydrops and/or with diagnosis of neonatal sepsis of uncertain aetiology was unavailable/difficult to analyse consistently even with text mining. Following additional adjustments were made to the point scoring of the biobank population. "Age, $\leq$ 18 years" (at diagnosis) was not applicable in the assessment of adult subjects. Information regarding the level of spleen enlargement and the presence of GD within relatives were unavailable. Plasma ferritin levels were used instead of serum ferritin. Information on disturbed motor function was difficult to analyse consistently. Information regarding the level of hepatomegaly, whether mild or moderate (originally both two points), or severe (one point) was not available and therefore all individuals with hepatomegaly were assigned two points.

2.4. Role of the funding source

The representatives of the funding source participated in study concept/design, interpretation of data, writing of the manuscript, and the decision to submit the paper for publication. All authors had access to the data of the study, within the limits of the General Data Protection Regulation and the Finnish Biobank Law. Only the personnel of the participating biobanks and Centogene had full access to the patient data and detailed *GBA* sequencing data (pseudonymised), respectively, in the context of this study. Corresponding author had final responsibility for the decision to submit for publication.

3. Results

3.1. Confirmation of the diagnosis of GD patients by GBA sequencing and lyso-Gb1 measurements

The five GD patients included in this study had originally been diagnosed with GD based on their clinical presentation, organ involvement, and available diagnostic laboratory analyses, including GlcCerase activity assay. The patients are recorded with the ICD-10 code E75.2 in the EHR of the Oulu University Hospital. According to the clinical evaluation, the patients represent type 1 GD (personal communication, Prof. Markku Savolainen, Oulu University Hospital, Oulu, Finland). In the current study, the patients were analysed for *GBA* variant status, and additionally, for lyso-Gb1 levels to confirm their diagnosis using updated molecular diagnostics and to test the feasibility of utilising lyso-Gb1 assay as a biomarker analysis.

All five patients had known pathogenic *GBA* mutations (Table 1). Patients 1 and 6 were homozygous while remaining three patients were compound heterozygotes for p.Asn409Ser (N370S). Patients 2 and 7 had p.Asn409Ser/p.Leu483Pro (N370S/L444P) and p.Asn409Ser/IVS2 + 1G > A genotypes, respectively (Table 1). Patient 5 harboured a novel deletion variant in the second allele, c.863delT, which results in a frameshift in translation (p.Asn409/p.Leu288fs genotype) (Table 1).

Although four out of five patients were on enzyme replacement therapy (ERT) at the time of sample donation, lyso-Gb1 levels were above the assay-specific thresholds in all study subjects (Table 2) which is line with the data in the literature [27]. Some variation was observed between the results obtained from the plasma and DBS samples with DBS samples showing 1.2–2.2 times higher lyso-Gb1 concentrations compared with frozen plasma samples (Table 2). Together with the *GBA* mutation status, these results confirm the molecular diagnosis of the previously diagnosed GD patients included in this study.

3.2. GED-C point scoring of GD patients

The GED-C PSS, with local adjustments (see 2.3 Point scoring, and Table A.1), was evaluated in five Finnish GD patients utilising longitudinal, retrospective EHR data before the treatment. The available EHR data allowed the retrospective assessment of 30 of the 32 original GED-C PSS signs/co-variables. Overall, point scores of all assessed patients were high, although variable, among inspected individuals (6–18.5 points based on the available data on 28–29 signs/co-variables per patient) (Table 3). According to the point distribution per sign/co-variable, the most prevalent sign/co-variables were mild or moderate anaemia (5/5 individuals), bone issues (4/5), family history of GD (4/5), splenomegaly (3/5), thrombocytopenia (3/5), leukopenia (3/5), and adult gammopathy (2/4) (Table 3). These results provide a preliminary range for the GED-C point score of Finnish GD patients.

3.3. GED-C point scoring and a subgroup with splenomegaly of biobank population

To explore the potential of retrospectively screening EHR data in

Table 1
GBA variant status of the five previously diagnosed Finnish Gaucher disease patients included in the study.

ID*	Genomic position of allele 1[a]	Predicted protein change of allele 1[b]	Genomic position of allele 2[a]	Predicted protein change of allele 2[b]	Result evaluation
GD_1	c.1226A > G	p.Asn409Ser (N370S)	c.1226A > G	p.Asn409Ser (N370S)	Affected
GD_2	c.1226A > G	p.Asn409Ser (N370S)	c.1448 T > C	p.Leu483Pro (L444P)	Affected
GD_5	c.1226A > G	p.Asn409Ser (N370S)	c.863delT	p.Leu288fs	Affected
GD_6	c.1226A > G	p.Asn409Ser (N370S)	c.1226A > G	p.Asn409Ser (N370S)	Affected
GD_7	c.1226A > G	p.Asn409Ser (N370S)	c.115 + 1G > A	- (IVS2 + 1G > A)	Affected

Abbreviations: GD, Gaucher disease.

* Patients recorded with the ICD-10 code E75.2.

[a] NCBI reference sequence: NM_000157.3. DNA sequences were not submitted to GenBank (for more information, see 2.2 GBA sequencing and lyso-Gb1 measurement).

[b] Nomenclature according to the recommendation of the Human Genome Variation Society (the first residue of the 39-residue signal sequence considered as number 1). The nomenclature often used in the literature in parenthesis (the first residue of the mature protein considered as number 1).

Table 2
Lyso-Gb1 levels determined from dried blood spots and plasma samples of the five previously diagnosed Finnish Gaucher disease patients included in the study.

ID*	ERT	Lyso-Gb1 (ng/ml)	
		DBS[a]	Plasma[b]
GD_1	Yes	10.3	5.5
GD_2	Yes	19.6	16.2
GD_5	Yes	28.2	13.1
GD_6	No	79.9	43.3
GD_7	Yes	107.0	62.8

Abbreviations: DBS, dry blood spots; ERT, enzyme replacement therapy; GD, Gaucher disease.

* Patients recorded with the ICD-10 code E75.2.

[a] A reference cut-off value for the lyso-Gb1 measurement from DBS, 6.8 ng/ml.

[b] A reference cut-off value for the lyso-Gb1 measurement from plasma, 1.2 ng/ml.

order to identify potential undiagnosed GD patients for diagnostic testing, the GED-C PSS was applied to EHR data representing a base population of 170,000 individuals previously treated at the special health care of the HDSF and who had donated samples to Auria Biobank. In this cohort, cases with pre-existing GD diagnosis (ICD-10 code E75.2) were not found. However, according to the general estimation on the worldwide prevalence of GD, this cohort likely contains several undiagnosed cases.

The available EHR data allowed the scoring of altogether 161,950 adult biobank subjects with 27 GED-C signs/co-variables (see 2.3 Point scoring, and Table A.1). Most of the assessed subjects had zero (n = 46,198, 28.5%) or two points (n = 62,168, 38.4%) (Fig. 1). Altogether 1,158 patients (0.72%) and 250 patients (0.15%) were assigned a score of ≥ 6 and ≥ 7.5 points, respectively, with highest point score observed being 13.5 points (n = 2, 0.001%) (Fig. 1).

All high-score individuals were subjected to lyso-Gb1 measurements to confirm or exclude GD at molecular level. However, only forty-two (16.8%) of individuals with a point score of ≥ 7.5 points (n = 250) had plasma samples at the biobank and could be tested so far. All analysed plasma samples were below the reference cut-off value for normal

Table 3
Overall GED-C point scores and distribution of points per sign/co-variable among the five previously diagnosed Finnish Gaucher disease patients included in the study.

ID*		GD_1		GD_2		GD_5		GD_6		GD_7	
Time from diagnosis (~years)		**2**		**25**		**21**		**1**		**19**	
ERT treatment started (years from diagnosis)		**0.5**		**17.6**		**0.6**		**No**		**0.16**	
Age at diagnosis		**51**		**18**		**5**		**50**		**5**	
Sex		**Female**		**Female**		**Female**		**Female**		**Male**	
Overall point scores		**13**		**18.5**		**14.5**		**6**		**12**	
Assessed GED-C PSS signs/co-variables[a]		**Yes / No**	**Points[b]**	**Yes / No**	**Points[b]**	**Yes / No**	**Points[b]**	**Yes / No**	**Points[b]**	**Yes / No**	**Points[b]**
3	Splenomegaly (≥3-fold enlargement)	No		Yes	3	Yes	3	No		Yes	3
	Disturbed oculomotor function (slow horizontal saccades with unimpaired vision)	Yes	3	No		No		No		No	
2	Thrombocytopenia, mild or moderate (platelet count, 50–150 × 10^9/l)	No		Yes	2	Yes	2	No		Yes	2
	Bone issues, including pain, crises, avascular necrosis and fractures	Yes	2	Yes	2	Yes	2	No		Yes	2
	Family history of Gaucher disease	Yes	2	Yes	2	Yes	2	Yes	2	No	
	Anaemia, mild or moderate (Hb, 95–140 g/l)	Yes	2	Yes	2	Yes	2	Yes	2	Yes	2
	Hyperferritinaemia, mild or moderate (serum ferritin, 300–1,000 μg/l)	No		Yes	2	No		Yes	2	No	
	Disturbed motor function (impairment primary motor development)	Yes	2	No		No		No		No	
	Hepatomegaly, mild or moderate (≤3-fold enlargement)	No		Yes	2	Yes	2	No		No	
	Myoclonus epilepsy	No		No		No		No		No	
	Kyphosis	No		No		No		No		Yes	2
	Adult gammopathy – monoclonal or polyclonal	Yes	2	Yes	2	No		No		NA	
1	Anaemia, severe (Hb, <95 g/l)	No		No		No		No		No	
	Hyperferritinaemia, severe (serum ferritin, >1,000 μg/l)	No		No		No		No		No	
	Hepatomegaly, severe (>3-fold enlargement)	No		No		No		No		No	
	Thrombocytopenia, severe (platelet count, <50 × 10^9/l)	No		No		No		No		No	
0.5	Gallstones	No		Yes	0.5	No		No		No	
	Bleeding, bruising or coagulopathy	No		No		Yes	0.5	No		No	
	Leukopenia	No		Yes	0.5	Yes	0.5	No		Yes	0.5
	Cognitive deficit	No		No		No		No		No	
	Low bone mineral density	No		No		No		No		No	
	Growth retardation including low body weight	No		No		No		No		No	
	Asthenia	No		No		No		No		No	
	Cardiovascular calcification	No		No		No		No		No	
	Dyslipidaemia	No		No		No		No		No	
	Elevated ACE levels	No		Yes	0.5	NA		NA		No	
	Fatigue	No		No		No		No		No	
	Pulmonary infiltrates	No		No		No		No		No	
	Age, ≤18 years	No		No		Yes	0.5	No		Yes	0.5
	Family history of Parkinson's disease	NA		NA		NA		No		No	

Abbreviations: ACE, angiotensin converting enzyme; ERT, enzyme replacement therapy; GED-C, Gaucher Earlier Diagnosis Consensus initiative; Hb, haemoglobin; NA, not assessed; PSS, point-scoring system.

* Patients recorded with the ICD-10 code E75.2.

[a] Information regarding Jewish ancestry and blood relative who died of foetal hydrops and/or with diagnosis of neonatal sepsis of uncertain aetiology, included in the original GED-C PSS, were not available (see 2.3 Point scoring, and Table A.1).

[b] Maximum number of points was 0.5–3 per sign/co-variable.

lyso-Gb1 level (Table A.3). The extremely low lyso-Gb1 levels are in line with the data previously reported on healthy controls [6,27].

Tissue samples were available for the remaining high-score individuals. However, formalin-fixed, paraffin-embedded tissue samples cannot be utilised in the lyso-Gb1 assay and their performance in genetic analyses is currently limited. Therefore, confirmation/exclusion of GD among the assessed biobank population warrants further studies.

Furthermore, the number of high-score individuals among assessed biobank population was rather high. Therefore, further prioritisation to diagnostic testing is needed. Splenomegaly, one of the most prevalent findings in GD and assigned 3 points in the GED-C PSS, was recorded for 408 individuals (0.25%) among the biobank population. These individuals were further subgrouped into six clusters according to the pattern of scored symptoms. The clusters were then explored in terms of point-score distribution and co-occurring ICD-10 diagnoses to assess whether any pattern of scored symptoms or other diagnoses associated with the high scores (Fig. A.1A). The point-score distributions were different between the clusters with cluster three having lower mean point score (3.8) and cluster four and five higher (8.9 and 8.3, respectively) than remaining clusters (Fig. A.1B and Table A.2). Neoplasms, diseases of the digestive system, and diseases of the circulatory system were the most prevalent across the clusters. Low-score cluster three had almost exclusively lower occurrence of diagnoses in each chapter compared with other clusters while highest occurrence of diagnoses was often observed in high-score clusters, i.e., in cluster one or alternatively in clusters four or five (Fig. A.1C).

This study demonstrates the feasibility of applying the prototype GED-C PSS to the Finnish EHR data by data mining but highlights a need for further prioritisation among high-score individuals to be able to select subpopulations for diagnostic testing.

4. Discussion

The GED-C PSS for GD [23] is a prototype and needs to be validated in confirmed GD patients. Recently, the GED-C PSS was validated in UK with 25 patients [24]. The objective of the present study was to test the GED-C PSS in Finnish GD patients by exploring retrospective EHR data representing the period before start of the treatment of GD. Consequently, the results obtained in Finnish patients suggest an indicative point-score range, 6–18.5, for confirmed GD patients (based on the available data on 28–29 signs/co-variables per patient). The data is roughly in line with the recent observations from UK [24], except that the score range obtained in the current study also covers potentially mild cases. The assessed patients included one male and four females with type 1 GD. The patients had variable age at diagnosis (5–51 years) and the *GBA* mutation status (four different mutations in different allelic combinations).

Only five patients were assessed with altogether four different pathogenic/potentially pathogenic *GBA* mutations, namely p.Asn409Ser (N370S), p.Leu483Pro (L444P), IVS2 + 1G > A9, and a potentially novel, p.Leu288fs. Therefore, it is not possible to make definitive conclusions on the genotype-phenotype correlation from these patients. However, it is known that p.Asn409Ser (N370S) mutation gives rise to most of the type 1 GD cases while p.Leu483Pro (L444P), especially if in homozygous form, is more common in types 2 and 3 [3,8,9,12,22]. Two patients homozygous for p.Asn409Ser (N370S) had received their diagnosis at 51 and 50 years of age, respectively, and had no recording on splenomegaly. It is possible that these patients may represent a mild

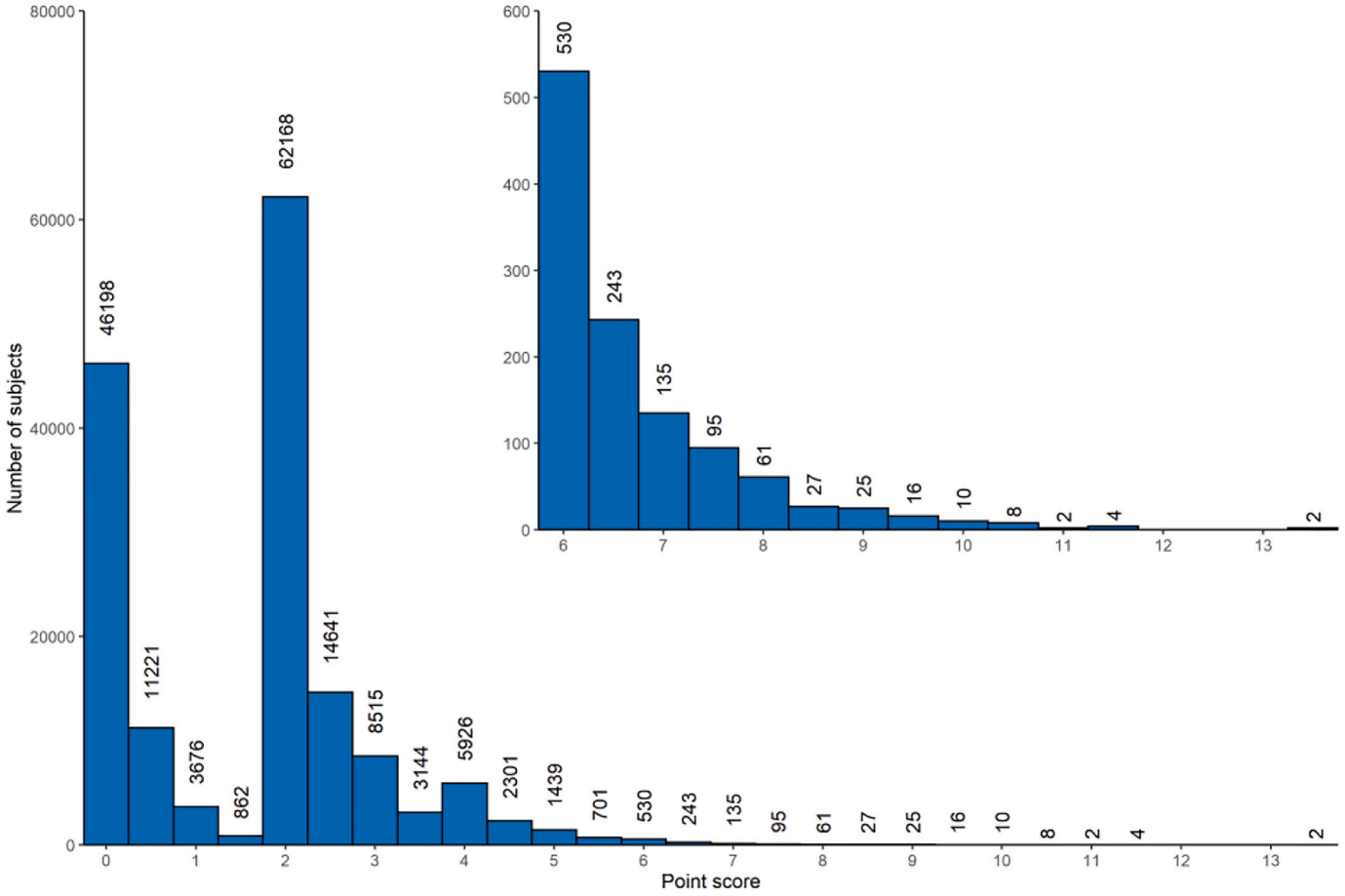

Fig. 1. GED-C point-score distribution in the biobank population of the Hospital District of Southwest Finland. The figure shows the point-score distribution among all assessed adult subjects ($N = 161{,}950$). In the insert, the distribution among subjects with 6 points or more is magnified.

phenotype associated with this mutation status. However, one of these patients had been recorded with disturbed oculomotor function, although p.Asn409Ser (N370S) in its homozygous form almost exclusively associates with type 1 without neurological symptoms [8,9]. Remaining patients were younger at diagnosis (5, 5, and 18 years). It also remains to be seen how common the potential novel pathogenic mutation, p.Leu288fs, is worldwide and what clinical presentations are associated with it. In the present study, one patient who was heterozygous for the p.Leu288fs mutation and was diagnosed for GD at five years of age, had the second highest point score among assessed subjects and was diagnosed with splenomegaly. Notably, only three of five GD patients were recorded with splenomegaly in this study. To determine whether splenomegaly is the most prevalent sign among Finnish GD patients, more patients should be evaluated.

There are currently approximately only 17 GD patients in Finland (personal communication, Prof. Markku Savolainen, Oulu University Hospital, Oulu, Finland), although the overall worldwide estimate on the prevalence of GD (1:30,000–100,000) suggests that 60–180 patients may potentially exist. Therefore, it is likely that most of the GD patients would have remained unidentified also in Finland. Accordingly, the second objective of this study was to test the feasibility of utilising GED-C PSS and Finnish EHR data for the screening of potential undiagnosed GD cases by data mining. EHR data was accessible via Auria Biobank, the first biobank in Finland. Finnish biobanks have extensive sample and data collections each representing hundreds of thousands of patients treated at one or more hospital districts. At the time of this study, Auria had around 170,000 biobank sample donors and EHR data collected at the Hospital District of Southwest Finland (HDSF). Longitudinal, retrospective EHR data was available from 161,950 adult subjects and allowed the screening for 27 of 32 original GED-C signs and co-variables. Only “family history of Gaucher disease”, “Jewish ancestry”, “disturbed motor function”, “blood relative who died of foetal hydrops and/or with diagnosis of neonatal sepsis of uncertain aetiology”, and “age, $\leq$ 18 years” (at diagnosis) had to be excluded. The data was either not available, did not allow consistent medical chart text mining across the cohort, or was not feasible due to the age of assessed biobank subjects.

This study is based on existing EHR data and thereby limited by the availability of the data to be collected. It is possible that in the data extraction process, information for all remaining 27 data fields was not available from all assessed subjects. The overall point-score distribution among biobank population showed that most of the screened individuals had zero or low number of points, while a subgroup of 1,158 individuals were assigned at least a score of 6 points, i.e., the lowest score observed among the confirmed GD patients included in this study. However, none of these “GD suspects” had a point score as high as 18.5. Confirmed GD patients were scored including three additional signs (“family history”, “disturbed motor function”, and “age, $\leq$ 18 years”) and if these signs would have been excluded, the total score range among confirmed GD cases would have been 4–16.5 points. This suggests that GD patients may be found also among the lower-score subjects of the biobank population.

Prioritisation of high-sore “GD suspects” for diagnostic testing to confirm the presence or absence of GD is challenging as the most prevalent sign of GD can be e.g., a mild or moderate anaemia, a relatively common condition. Splenomegaly was less prevalent and recorded in only 0.25% of biobank subjects but was associated with rather variable point scores and ICD-10 diagnoses.

In the current study, a substantial amount of high-score biobank subjects remained to be tested and the type of available archived samples will dictate the methods to be utilised in subsequent diagnostic studies. It is also possible that undiagnosed GD patients don’t exist in the assessed biobank population due to the fact that the prevalence of GD, an inherited rare disorder, might be highly variable in different parts of the country.

Therefore, additional biobank populations and further characterisation of “GD suspects” with better algorithms and data-mining tools accompanied by large-scale testing of biobank samples with alternative approaches to diagnostic testing are needed to address to these challenges in the future.

Contributors

Conceptualisation: Kaisa Elomaa, Carlos Vaca Falconi. Methodology: All authors. Validation: Markku J. Savolainen and Samppa Rohkimainen performed the GED-C point scoring of confirmed Gaucher patients. Antti Karlsson carried out the GED-C point scoring and the point-score distribution and cluster analyses of the biobank population. Aggregate data results (pseudonymised) were validated by all authors. Formal analysis: Antti Karlsson, Iiro Toppila. Investigation: Markku J. Savolainen, Samppa Rohkimainen, Antti Karlsson, and the personnel of the participating biobanks. Resources: Participating institutions/parties. Funding: Funding from Shire (now part of Takeda). Data Curation: Participating biobanks and Medaffcon. Writing - Original Draft: Kristiina Uusi-Rauva. Writing - Review & Editing: All authors. Visualisation: Design and drawing of figures by Antti Karlsson, Iiro Toppila, Kristiina Uusi-Rauva. Design and editing of data tables by all authors.

Declaration of Competing Interest

CVF and KE are employed by Takeda (Stockholm, Sweden and Helsinki, Finland, respectively). AK and SR are employed by Auria Biobank and Biobank Borealis, respectively, which received reimbursement from Shire (now part of Takeda), for the work done at Auria/Borealis. IT, KU, and MIL are employed by Medaffcon Oy (Espoo, Finland) which received funding from Shire (now part of Takeda), for conducting the study. MJS reports personal consultancy fees and travel grants from Shire (now part of Takeda) during the study, as well as grant support, paid to his institution, from several foundations for research outside the submitted work.

Acknowledgements

This study benefited from the samples/data from the Auria Biobank, Turku, Finland (https://www.auria.fi/biopankki/en/index.php?lang=en), and the Biobank Borealis of Northern Finland, Oulu, Finland (https://www.ppshp.fi/Tutkimus-ja-opetus/Biopankki/Pages/default.aspx). The personnel of the biobanks are thanked for their valuable help. The pathologist of the Biobank Borealis, MD PhD Saila Kauppila, is thanked for her assistance with specimen availability and histological confirmation. PhD Emma-Leena Alarmo (Takeda) is thanked for critical review of the manuscript. PhD Maija Wolf (previously employed by Medaffcon) is acknowledged for study management. The study was funded by Shire (now part of Takeda).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ymgmr.2021.100725.

References

[1] R.O. Brady, J.N. Kanfer, D. Shapiro, Metabolism of glucocerebrosides II. Evidence of an enzymatic deficiency in Gaucher’s disease, Biochem. Biophys. Res. Commun. 18 (1965) 221–225, https://doi.org/10.1016/0006-291X(65)90743-6.

[2] A. Dandana, S. Ben Khelifa, H. Chahed, A. Miled, S. Ferchichi, Gaucher disease: clinical, biological and therapeutic aspects, Pathobiology 83 (2016) 13–23, https://doi.org/10.1159/000440865.

[3] J. Stirnemann, N. Belmatoug, F. Camou, C. Serratrice, R. Froissart, C. Caillaud, T. Levade, L. Astudillo, J. Serratrice, A. Brassier, C. Rose, T. Billette de Villemeur, M. Berger, A review of Gaucher disease pathophysiology, clinical presentation and treatments, IJMS 18 (2017) 441, https://doi.org/10.3390/ijms18020441.

[4] J.M.F.G. Aerts, W.W. Kallemeijn, W. Wegdam, M. Joao Ferraz, M.J. van Breemen, N. Dekker, G. Kramer, B.J. Poorthuis, J.E.M. Groener, J. Cox-Brinkman, S. M. Rombach, C.E.M. Hollak, G.E. Linthorst, M.D. Witte, H. Gold, G.A. van der Marel, H.S. Overkleeft, R.G. Boot, Biomarkers in the diagnosis of lysosomal storage

disorders: proteins, lipids, and inhibodies, J. Inherit. Metab. Dis. 34 (2011) 605–619, https://doi.org/10.1007/s10545-011-9308-6.
[5] O. Nilsson, J.-E. Månsson, G. Håkänsson, L. Svennerholm, The occurrence of psychosine and other glycolipids in spleen and liver from the three major types of gaucher's disease, Biochimica et Biophysica Acta (BBA) - Lipids and Lipid Metabolism 712 (1982) 453–463, https://doi.org/10.1016/0005-2760(82)90272-7.
[6] A. Rolfs, A.-K. Giese, U. Grittner, D. Mascher, D. Elstein, A. Zimran, T. Böttcher, J. Lukas, R. Hübner, U. Gölnitz, A. Röhle, A. Dudesek, W. Meyer, M. Wittstock, H. Mascher, Glucosylsphingosine is a highly sensitive and specific biomarker for primary diagnostic and follow-up monitoring in Gaucher disease in a non-Jewish, Caucasian cohort of Gaucher disease patients, PLoS One 8 (2013), e79732, https://doi.org/10.1371/journal.pone.0079732.
[7] J. Charrow, H.C. Andersson, P. Kaplan, E.H. Kolodny, P. Mistry, G. Pastores, B. E. Rosenbloom, C.R. Scott, R.S. Wappner, N.J. Weinreb, A. Zimran, The Gaucher registry: demographics and disease characteristics of 1698 patients with Gaucher disease, Arch. Intern. Med. 160 (2000) 2835, https://doi.org/10.1001/archinte.160.18.2835.
[8] K.S. Hruska, M.E. LaMarca, C.R. Scott, E. Sidransky, Gaucher disease: mutation and polymorphism spectrum in the glucocerebrosidase gene (GBA), Hum. Mutat. 29 (2008) 567–583, https://doi.org/10.1002/humu.20676.
[9] V. Koprivica, D.L. Stone, J.K. Park, M. Callahan, A. Frisch, I.J. Cohen, N. Tayebi, E. Sidransky, Analysis and classification of 304 mutant alleles in patients with type 1 and type 3 Gaucher disease, Am. J. Hum. Genet. 66 (2000) 1777–1786, https://doi.org/10.1086/302925.
[10] L. Smith, S. Mullin, A.H.V. Schapira, Insights into the structural biology of Gaucher disease, Exp. Neurol. 298 (2017) 180–190, https://doi.org/10.1016/j.expneurol.2017.09.010.
[11] A. Mehta, N. Belmatoug, B. Bembi, P. Deegan, D. Elstein, Ö. Göker-Alpan, E. Lukina, E. Mengel, K. Nakamura, G.M. Pastores, J. Pérez-López, I. Schwartz, C. Serratrice, J. Szer, A. Zimran, M. Di Rocco, Z. Panahloo, D.J. Kuter, D. Hughes, Exploring the patient journey to diagnosis of Gaucher disease from the perspective of 212 patients with Gaucher disease and 16 Gaucher expert physicians, Mol. Genet. Metab. 122 (2017) 122–129, https://doi.org/10.1016/j.ymgme.2017.08.002.
[12] G.A. Grabowski, Phenotype, diagnosis, and treatment of Gaucher's disease, Lancet 372 (2008) 1263–1271, https://doi.org/10.1016/S0140-6736(08)61522-6.
[13] M.R. Alaei, A. Tabrizi, N. Jafari, H. Mozafari, Gaucher disease: new expanded classification emphasizing neurological features, Iran J Child Neurol. 13 (2019) 7–24.
[14] M.J. Eblan, O. Goker-Alpan, E. Sidransky, Perinatal lethal GAUCHER disease: a distinct phenotype along the NEURONOPATHIC continuum, Fetal and Pediatric Pathology. 24 (2005) 205–222, https://doi.org/10.1080/15227950500405296.
[15] O. Goker-Alpan, R. Schiffmann, J.K. Park, B.K. Stubblefield, N. Tayebi, E. Sidransky, Phenotypic continuum in neuronopathic gaucher disease: an intermediate phenotype between type 2 and type 3, J. Pediatr. 143 (2003) 273–276, https://doi.org/10.1067/S0022-3476(03)00302-0.
[16] G.A. Grabowski, A. Zimran, H. Ida, Gaucher disease types 1 and 3: phenotypic characterization of large populations from the ICGG Gaucher registry: phenotypes of Gaucher disease types 1 and 3, Am. J. Hematol. 90 (2015) S12–S18, https://doi.org/10.1002/ajh.24063.
[17] P. Dubot, L. Astudillo, N. Therville, F. Sabourdy, J. Stirnemann, T. Levade, N. Andrieu-Abadie, Are glucosylceramide-related sphingolipids involved in the increased risk for Cancer in Gaucher disease patients? Review and hypotheses, Cancers 12 (2020) 475, https://doi.org/10.3390/cancers12020475.
[18] B.E. Rosenbloom, N.J. Weinreb, A. Zimran, K.A. Kacena, J. Charrow, E. Ward, Gaucher disease and cancer incidence: a study from the Gaucher registry, Blood 105 (2005) 4569–4572, https://doi.org/10.1182/blood-2004-12-4672.
[19] A. Migdalska-Richards, A.H.V. Schapira, The relationship between glucocerebrosidase mutations and Parkinson disease, J. Neurochem. 139 (2016) 77–90, https://doi.org/10.1111/jnc.13385.
[20] A. Mehta, Epidemiology and natural history of Gaucher's disease, European Journal of Internal Medicine. 17 (2006) S2–S5, https://doi.org/10.1016/j.ejim.2006.07.005.
[21] P.J. Meikle, Prevalence of lysosomal storage disorders, JAMA 281 (1999) 249, https://doi.org/10.1001/jama.281.3.249.
[22] J. Stirnemann, M. Vigan, D. Hamroun, D. Heraoui, L. Rossi-Semerano, M.G. Berger, C. Rose, F. Camou, C. de Roux-Serratrice, B. Grosbois, P. Kaminsky, A. Robert, C. Caillaud, R. Froissart, T. Levade, A. Masseau, C. Mignot, F. Sedel, D. Dobbelaere, M.T. Vanier, V. Valayanopoulos, O. Fain, B. Fantin, T. de Villemeur, F. Mentré, N. Belmatoug, The French Gaucher's disease registry: clinical characteristics, complications and treatment of 562 patients, Orphanet J Rare Dis. 7 (2012) 77, https://doi.org/10.1186/1750-1172-7-77.
[23] A. Mehta, D.J. Kuter, S.S. Salek, N. Belmatoug, B. Bembi, J. Bright, S. vom Dahl, F. Deodato, M. Di Rocco, O. Göker-Alpan, D.A. Hughes, E.A. Lukina, M. Machaczka, E. Mengel, A. Nagral, K. Nakamura, A. Narita, B. Oliveri, G. Pastores, J. Pérez-López, U. Ramaswami, I.V. Schwartz, J. Szer, N.J. Weinreb, A. Zimran, Presenting signs and patient co-variables in Gaucher disease: outcome of the Gaucher Earlier Diagnosis Consensus (GED-C) Delphi initiative, Intern. Med. J. 49 (2019) 578–591, https://doi.org/10.1111/imj.14156.
[24] A. Mehta, O. Rivero-Arias, M. Abdelwahab, S. Campbell, A. McMillan, M.J. Rolfe, J.R. Bright, D.J. Kuter, Scoring system to facilitate diagnosis of Gaucher disease, Intern. Med. J. 50 (2020) 1538–1546, https://doi.org/10.1111/imj.14942.
[25] P.J.V. Seeman, Finckh U. MD, J. Hoppner, V. Lakner, I. Liebisch, G. Grau, A. Rolfs, Two new missense mutations in a non-Jewish Caucasian family with type 3 Gaucher disease, Neurology 46 (1996) 1102–1107, https://doi.org/10.1212/WNL.46.4.1102.
[26] U. Finckh, P. Seeman, O.C. von Widdern, A. Rolfs, Simple PCR amplification of the entire Glucocerebrosidase gene (GBA) coding region for diagnostic sequence analysis, DNA Seq. 8 (1998) 349–356, https://doi.org/10.3109/10425179809020896.
[27] S. Revel-Vilk, M. Fuller, A. Zimran, Value of Glucosylsphingosine (Lyso-Gb1) as a biomarker in Gaucher disease: a systematic literature review, IJMS 21 (2020) 7159, https://doi.org/10.3390/ijms21197159.