Molecular Genetics and Metabolism Reports 33 (2022) 100911

Contents lists available at ScienceDirect

Molecular Genetics and Metabolism Reports

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

Screening for potential undiagnosed Gaucher disease patients: Utilisation of the Gaucher earlier diagnosis consensus point-scoring system (GED-C PSS) in conjunction with electronic health record data, tissue specimens, and small nucleotide polymorphism (SNP) genotype data available in Finnish biobanks

Minja Pehrsson [a,1], Hanna Heikkinen [a,1], Ulla Wartiovaara-Kautto [b], Sampo Mäntylahti [a], Pia Bäckström [a], Mariann I. Lassenius [c], Kristiina Uusi-Rauva [c], Olli Carpén [a,d], Kaisa Elomaa [e,*]

[a] Helsinki Biobank, Helsinki University Hospital, Haartmaninkatu 3, 00290 Helsinki, Finland
[b] Department of Hematology, Comprehensive Cancer Center, Helsinki University Hospital and University of Helsinki, Haartmaninkatu 4, 00290 Helsinki, Finland
[c] Medaffcon Oy, Metsänneidonkuja 8, 02130 Espoo, Finland
[d] Department of Pathology, University of Helsinki, HUS Diagnostic Center, Finland
[e] Takeda Oy, Ilmalantori 1, 00101 Helsinki, Finland

ARTICLE INFO

Keywords:
Biobank study
Electronic health record data
Small nucleotide polymorphism chip genotype data
Gaucher disease
Gaucher earlier diagnosis consensus point-scoring system
GBA

ABSTRACT

Background: Autosomal recessive Gaucher disease (GD) is likely underdiagnosed in many countries. Because the number of diagnosed GD patients in Finland is relatively low, and the true prevalence is currently not known, it was hypothesized that undiagnosed GD patients may exist in Finland. Our previous study demonstrated the applicability of Gaucher Earlier Diagnosis Consensus point-scoring system (GED-C PSS; Mehta et al., 2019) and Finnish biobank data and specimens in the automated point scoring of large populations. An indicative point-score range for Finnish GD patients was determined, but undiagnosed patients were not identified partly due to high number of high-score subjects in combination with a lack of suitable samples for diagnostics in the assessed biobank population. The current study extended the screening to another biobank and evaluated the feasibility of utilising the automated GED-C PSS in conjunction with small nucleotide polymorphism (SNP) chip genotype data from the FinnGen study of biobank sample donors in the identification of undiagnosed GD patients in Finland. Furthermore, the applicability of FFPE tissues and DNA restoration in the next-generation sequencing (NGS) of the *GBA* gene were tested.

Methods: Previously diagnosed Finnish GD patients eligible to the study, and up to 45,100 sample donors in Helsinki Biobank (HBB) were point scored. The GED-C point scoring, adjusted to local data, was automated, but also partly manually verified for GD patients. The SNP chip genotype data for rare *GBA* variants was visually assessed. FFPE tissues of GD patients were obtained from HBB and Biobank Borealis of Northern Finland (BB).

Results: Three previously diagnosed GD patients and one patient previously treated for GD-related features were included. A genetic diagnosis was confirmed for the patient treated for GD-related features. The GED-C point score of the GD patients was 12.5–22.5 in the current study. The score in eight Finnish GD patients of the previous and the current study is thus 6–22.5 points per patient. In the automated point scoring of the HBB

Abbreviations: BB, Biobank Borealis of Northern Finland; DF4/DF5, Data freeze 4/5; EHR, Electronic health record; FFPE, Formalin-fixed, paraffin embedded; GBA1/GBA, β-glucocerebrosidase gene; GD, Gaucher disease; GlcCer, β-glucosylceramide; GlcCerase, β-glucosylceramidase; GlcSph/Lyso-Gb1, β-glucosyl-sphingosine; GED-C, Gaucher Earlier Diagnosis Consensus; HBB, Helsinki Biobank; HUH, Helsinki University Hospital; HUS, Hospital District of Helsinki and Uusimaa; ICD-10, International Statistical Classification of Diseases and Related Health Problems 10th Revision; NGS, Next-generation sequencing; OUH, Oulu University Hospital; PSS, Point-scoring system; SNP, small nucleotide polymorphism; VUS, variant of uncertain significance.

* Corresponding author at: Nordic Innovation CoE Lead, Takeda, PO Box 1406, Ilmalantori 1, 00101 Helsinki, Finland.
E-mail addresses: minja.pehrsson@hus.fi (M. Pehrsson), hanna.m.heikkinen@iki.fi (H. Heikkinen), ulla.wartiovaara-kautto@hus.fi (U. Wartiovaara-Kautto), sampo.mantylahti@hus.fi (S. Mäntylahti), pia.j.backstrom@hus.fi (P. Bäckström), mariann.lassenius@medaffcon.fi (M.I. Lassenius), kristiina.uusi-rauva@medaffcon.fi (K. Uusi-Rauva), olli.carpen@helsinki.fi (O. Carpén), kaisa.elomaa@takeda.com (K. Elomaa).
[1] These authors contributed equally.

https://doi.org/10.1016/j.ymgmr.2022.100911
Received 9 June 2022; Received in revised form 12 August 2022; Accepted 12 August 2022
2214-4269/© 2022 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

subpopulation (N ≈ 45,100), the overall scores ranged from 0 to 17.5, with 0.77% (346/45,100) of the subjects having ≥10 points. The analysis of SNP chip genotype data was able to identify the diagnosed GD patients, but potential undiagnosed patients with the GED-C score and/or the *GBA* genotype indicative of GD were not discovered. Restoration of the FFPE tissue DNA improved the quality of the *GBA* NGS, and pathogenic *GBA* variants were confirmed in five out of six unrestored and in all four restored FFPE DNA samples.
Discussion: These findings imply that the prevalence of diagnosed patients (~1:325,000) may indeed correspond the true prevalence of GD in Finland. The SNP chip genotype data is a valuable tool that complements the screening with the GED-C PSS, especially if the genotyping pipeline is tuned for rare variants. These proof-of-concept biobank tools can be adapted to other rare genetic diseases.

1. Introduction

Gaucher disease (GD), an autosomal recessive disorder caused by the deficiency of a lysosomal enzyme β-glucosylceramidase (GlcCerase; EC3.2.1.45), represents a rare disease that is treatable but likely underdiagnosed in many countries [1–3]. The true prevalence of GD in Finland has remained elusive. Finnish GD patients are typically being investigated for their symptoms in secondary or tertiary health care (central or university hospitals, respectively), and there are currently <20 diagnosed GD patients, thus corresponding to a prevalence of ~1:325,000 (U. Wartiovaara-Kautto, docent, acting chief physician, Department of Hematology, Helsinki University Hospital Comprehensive Cancer Centre; personal communication). This is lower than reported worldwide estimations and population-specific prevalences, that range from ~1:1,000 among Ashkenazi Jewish to 1:136,000 in France [4–6]. Therefore, it is possible that undiagnosed GD patients exist in Finland.

Availability of GD diagnostic laboratory analyses is good in Finland. The low diagnostic rate in GD may however result from a lack of general disease awareness, and a varying spectrum of symptoms that overlap with other conditions, thus further complicating the diagnostic process [7,8]. The wide phenotypic spectrum of GD at diagnosis includes e.g., thrombocytopenia, splenomegaly, hepatomegaly, bone or joint pain, and anaemia, the latter two, for example, representing relatively common conditions [7]. The time of onset, symptoms, and disease severity are utilised in disease subtyping (the subtypes 1–3) [3,9]. The diagnosis of GD is based on clinical presentation, organ involvement, and diagnostic laboratory analysis, including GlcCerase activity assay from blood or skin fibroblasts, biomarker analysis from blood (e.g., β-glucosylsphingosine; GlcSph/lyso-Gb1), histological examinations of cells with glucosylceramides (GlcCer) and other lipid deposits, and genetic tests of the GlcCerase-encoding *GBA* gene (glucosidase, beta, acid; MIM# 606463; also known as GBA1) [3]. Patients may have had appointments at different specialities before obtaining a diagnosis for GD, several years after first consulting a doctor [7,8,10].

To increase the diagnostic rate in GD, high-risk subjects, e.g., patients diagnosed with splenomegaly, thrombocytopenia, or plasma cell dyscrasias, have been screened by utilising standard enzyme, biomarker, and genetic diagnostic tools [11,12]. In addition, larger cohorts comprising of up to ca 5,300 individuals, have been screened by utilising left-over blood samples collected in routine care or available laboratory data a priori followed by the testing of separately collected specimens [13,14]. The prevalence of GD in the tested high-risk and hospital populations was 0–3.3% and 0–0.019%, respectively [11–14].

We have recently introduced a novel approach for the screening of undiagnosed GD patients in Finland [15]. If successful, the approach that utilises Gaucher Earlier Diagnosis Consensus point-scoring system (GED-C PSS) [16], automated point scoring of longitudinal electronic health record (EHR) data, and Finnish biobank data and specimens allows the screening of large populations. Finnish biobanks operate in conjunction with the major hospitals and hospital districts and represent a valuable data and sample source for research in accordance with the Finnish Biobank Act 688/2012. The biobank samples can be linked with respective EHR data and previously obtained sample-derived data, thus allowing comprehensive analyses. Moreover, Finnish biobank sample donors can be recontacted, if the donor has given a consent for such process and referred to clinical examination to adequately confirm/exclude potential findings.

The applicability of the Finnish EHR data of the secondary/tertiary health care and the GED-C PSS in the automated point scoring was demonstrated in our previous study assessing ca 160,000 sample donors in Auria biobank (Turku, Finland) [15]. The GED-C PSS of 32 signs/covariables, originally developed for patient chart review-based point scoring of type 1 and type 3 GD, independently from the current study, represents a prototype PSS that hasn't been assigned for a commonly accepted cut-off value but which has so far been validated in 25 UK patients as well as tested in five Finnish GD patients [15–17]. In the five Finnish GD type 1 patients, the GED-C point score was 6–18.5 per 28–29 manually assessed signs/covariables per patient [15]. In the screened Auria biobank population, the score range was 0–13.5 in the automated assessment of 27 GED-C signs/covariables with 0.72% (n ≈ 1,160) of the subjects having a score of six or more points [15]. Undiagnosed GD patients were not identified in a subsequent biomarker analysis. However, the number of high-score subjects was rather high in the tested biobank population, and the high-score subjects partly remained to be tested due to lack of plasma and blood DNA samples suitable for the analysis of GlcCerase activity, lyso-Gb1 biomarker levels, or *GBA* sequencing [15].

In addition to the blood samples, the Finnish biobanks host a vast number of research and diagnostic formalin-fixed and paraffin-embedded (FFPE) tissue specimens that could be utilised in e.g., next-generation sequencing (NGS). However, the performance of DNA extracted from potentially long-term stored FFPE tissue specimens with respect to high-quality blood DNA in the NGS of the *GBA* gene needs to be evaluated before any large-scale analysis. Moreover, GED-C point-scored high-score subjects could be further prioritised for diagnostic laboratory analysis by utilising the increasing volume of small nucleotide polymorphism (SNP) chip genotype data generated in the FinnGen study, an ongoing independent genome-wide association study of blood samples collected from Finnish biobanks [18].

In the current study, we have extended our biobank-based GD screening approach to another Finnish biobank, Helsinki Biobank (Helsinki, Finland; HBB) and aimed to obtain more data on the GED-C point-score range among Finnish GD patients in support of automated point scoring, to test the applicability of FFPE tissue specimens of GD patients in *GBA* NGS, and to evaluate the feasibility of utilising the automated GED-C PSS in conjunction with the SNP chip genotype data in the identification of potential undiagnosed GD patients for diagnostic testing in Finland.

2. Materials and methods

2.1. Ethics and study design

This retrospective biobank study, governed by the Finnish Biobank Act 688/2012, was conducted in collaboration with Helsinki Biobank (HBB; Helsinki, Finland) and Biobank Borealis of Northern Finland (BB; Oulu, Finland). Samples and respective data available in BB were only

utilised in the assessment of the quality of the FFPE tissue specimens in *GBA* NGS.

Requests for a study approval together with a detailed study protocol and a sample/data request were submitted to each biobank (respective biobank project numbers: HBP20200086 and BB_2021_5006) and to the ethics committee of the Hospital District of Helsinki and Uusimaa (HUS/2208/2020). Informed consents were not required as stated by the ethics committee. Inclusion of the FFPE tissue specimens from BB did not require a statement from the local ethics committee. The study was a non-interventional study limited to the use of readily available biobank specimens and associated pseudonymised specimen-derived and clinical data and did not involve recontacting the patients.

2.2. Eligibility and study cohorts

Throughout the current biobank study, only those biobank sample donors whose biobank sample(s) and/or sample-derived data were adequately and readily available for the analyses of the current study were included.

HBB sample donors were initially screened for the following records available by the end of February 2021 in the EHRs of the secondary/tertiary health care of the Hospital District of Helsinki and Uusimaa (HUS): structural records on the International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10) code E75* (Disorders of sphingolipid metabolism and other lipid storage disorders, including GD), structural records on GD-specific diagnostics and biomarker analyses, and other GD-related terms provided as free-format texts in medical charts. The initial screening was followed by an inspection of sample availability for the determination of study and cohort eligibility. HBB sample donors with data available from the FinnGen study were separately identified.

Final study cohorts and respective subgroups with the number of subjects are illustrated in Fig. 1. The cohort 1 ($N = 3$) included HBB sample donors who have been diagnosed with GD in Helsinki University Hospital (HUH), HUS. The cohort 2 ($N = 1$) consisted of HBB sample donors who had been treated in HUH and represented with potential GD-related features ("suspected GD patients"), i.e., patients who have been subjected to GD-specific diagnostic and/or biomarker test or who have other potential GD-related recording(s) (e.g., free-text entries) but no definitive records of diagnosis. The cohort 3 ($N = 7$) included the two previously diagnosed GD patients and the "suspected GD patient" of the cohorts 1 and 2, respectively, with FFPE tissue specimens available in

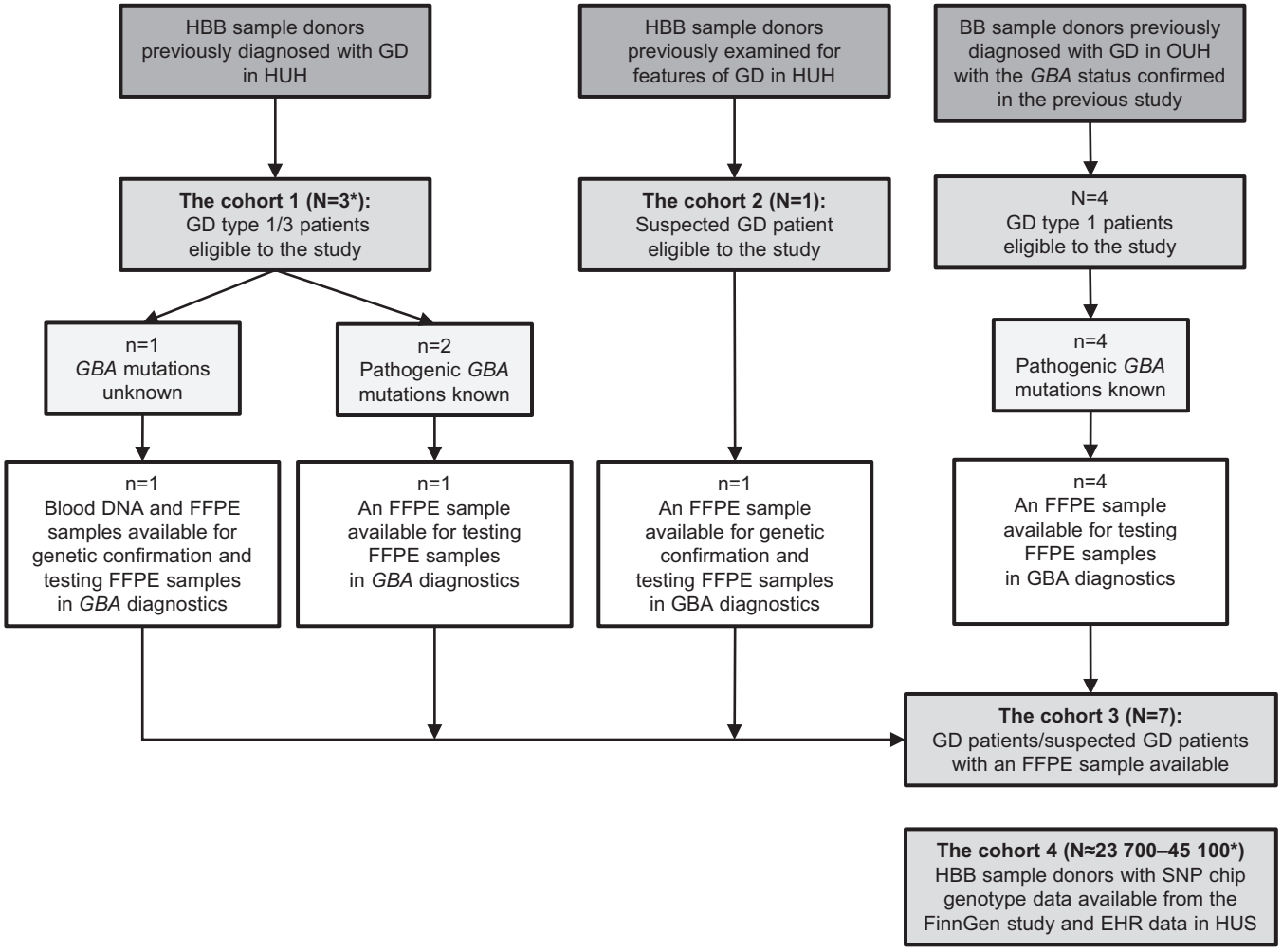

Fig. 1. A summary of the formation of the final cohorts 1–4. Subjects of the cohort 1 and 2 represented patients previously diagnosed with or examined for features of Gaucher disease (GD) in Helsinki University Hospital (HUH) and who had samples/sample-derived data available in Helsinki Biobank (HBB). In addition, patients with known *GBA* variant status [15] and diagnosed with GD in Oulu University Hospital (OUH) with formalin-fixed, paraffin-embedded (FFPE) tissue samples available in Biobank Borealis (BB) were included to increase the number of samples in the cohort 3. Samples of the subjects of the cohort 1 have been analysed in the FinnGen study and are thus included in the final cohort 4 with genotype data as well as electronic health record (EHR) data available in Hospital District of Helsinki and Uusimaa (HUS) (indicated by an asterisk).

HBB, and additionally, four GD patients previously diagnosed in Oulu University Hospital (OUH) with pathogenic *GBA* variant status confirmed in the previous study [15] and an FFPE tissue specimen available in BB. The cohort 4 (N ≈ 23,700–45,100) represented a subpopulation of HBB sample donors with SNP chip genotype data available from the FinnGen study and EHR data in the records of HUS. In the cohort 4, the following subgroups were formed based on *GBA* genotypes determined for a subset of sample donors by the cluster plot analysis of the SNP chip genotype data from the FinnGen study (for details and final number of subjects, see 3 Results): A) Previously diagnosed GD patients (the cohort 1) ($N = 3$; one homozygote and two compound heterozygotes for pathogenic *GBA* variants); B) Additional potential homozygotes/compound heterozygotes for pathogenic *GBA* variants ($N = 0$); C) Potential heterozygote carriers of pathogenic hot spot variants c.1448 T > C and c.1226A > G ($N = 55$); D) Controls likely negative for pathogenic hot spot variants c.1448 T > C and c.1226A > G ($N = 4{,}670$).

2.3. GED-C point scoring

The original GED-C PSS was proposed by the experts of the GED-C initiative independently and outside of the current study [16]. The GED-C point scoring was carried out for the cohorts 1, 2, and 4, and was based on longitudinal, retrospective EHR data of the secondary/tertiary health care of HUS. GED-C signs/covariables were applied with adequate local adjustments to match the local data format, Finnish language, and data availability (Table 1). In addition, two additional signs introduced in Mehta et al. [17] were separately assessed (splenectomy and lymph node enlargement). Information regarding the extent of splenomegaly or hepatomegaly was not consistently available. Hepatomegaly was thus always assigned two points. Furthermore, data for family history-related GED-C signs/covariables were not text mined due to unavailability of such data in general. Family history of GD in the point scoring of the cohorts 1 and 2 was manually reviewed from medical charts. For signs/covariables that were based on laboratory tests, the status and score were determined based on the most recent prevalent status.

The point scoring was automated, but also partly verified manually in a chart review, for the cohorts 1 and 2. In the assessment of the cohorts 1 and 2, only data representing time before the treatment of GD was considered. However, where the study cohorts 1/2 and 4 were compared, only signs/covariables enabling automated point scoring and data without time restrictions were included (Table 1, Automated point scoring). Therefore, the results of the automated assessment may differ from the ones obtained with the manual verification.

Scripts utilised in the point scoring are stored in HBB and cannot be exported out of the data analysis environment due to data protection regulations.

2.4. Analysis of GBA genotypes from the SNP chip genotype data generated in the FinnGen study

SNP chip genotype data was available from ca 47,600 HBB sample donors genotyped in the FinnGen study, independently from the current study [18]. The FinnGen is an ongoing study, and data is provided to HBB biannually. Therefore, Data freeze 4 SNP chip genotype data was primarily utilised in the current study (DF4; 34,200 HBB sample donors in total) with extensions to Data freeze 5 (DF5; 47,600 HBB sample donors in total) as indicated in the results.

Blood samples in the FinnGen were originally genotyped with Affymetrix arrays and SNP chip calling algorithms as previously described [19]. The FinnGen array chip allows the analysis of approximately 30 variants located in the *GBA* gene region with some of the variants being likely benign, of uncertain significance (VUS), or not reported in ClinVar. Only pathogenic/likely pathogenic/VUS were analysed in the study (Table 2). It should be noted that the analysed samples may include *GBA* variants that are not covered by the FinnGen array, including e.g., a hot spot variant c.115 + 1G > A, as well as c.681 T > G (harboured by one GD patient in the current study). Furthermore, the analysis of DF4 (chip v1) and DF5 (chip v2) data sets are based on slightly different sets of variants due to updates of the probe sets (Table 2).

The SNP chip calling algorithms are generally not reliable for genotyping very rare variants, and thus the computed genotype results may contain wrong positive/negative samples [21]. Therefore, the genotypes of pathogenic or likely pathogenic *GBA* variants, as reported in the ClinVar database (Table 2), were manually determined from cluster plots of the SNP chip genotype data for each variant of interest. The cluster plots were analysed using Axiom Analysis Suite software. Where indicated, only a subset of genotype data was assessed.

The formation of genotype-based subgroups was carried out as follows. Representative samples determined as no call (i.e., genotype was missed, or sample failed) or as initially heterozygote carriers for the *GBA* hot spot variants c.1448 T > C or c.1226A > G in the cluster plots were chosen for the validation by *GBA* NGS. Note that the cluster plot data of the remaining variants in Table 2 was not validated by sequencing. However, it has been reported that c.1448 T > C and c.1226A > G, together with c.84dupG and c.115 + 1G > A (IVS2 + 1G > A), constitute approximately 50–60% of known pathogenic *GBA* variants in non-Jewish populations [22]. Based on the cluster plot analysis, there were likely no samples with c.84dupG, and c.115 + 1G > A cannot be currently analysed with the FinnGen array chip. Therefore, validation was considered adequate. Samples confirmed positive for the c.1448 T > C or c.1226A > G variants in the validation sequencing were used as reference and the cluster plot positions of altogether 23,700 samples of 34,200 samples in total of the DF4 were analysed from the cluster plots for the formation of the genotype-based subgroups of the cohort 4 to be point scored.

In addition to the subset of samples analysed for the formation of the *GBA* genotype-based subgroups, the *GBA* genotypes of all GED-C point-scored high-score samples (the GED-C point score ≥ 10; $n = 346$) of DF5 (N ≈ 45,100) were analysed from the cluster plots as described above. The genotypes of the samples that were considered potential homozygotes/compound heterozygotes based on the cluster plot analysis were validated by *GBA* NGS throughout the study.

2.5. DNA processing

Blood-derived DNA samples that were sequenced in the current study were readily available in the biobank. FFPE tissue-derived DNA was extracted according to manufacturer's instructions (QS GeneRead™ DNA FFPE treatment kit and QS DSP™ DNA mini kit, respectively, Qiagen, Hilden, Germany) (see Appendix A). FFPE tissue-derived DNA samples were also processed for additional DNA restoration step performed according to manufacturer's instructions (Infinium HD FFPE QC Kit and Infinium HD FFPE Restore Kit, Illumina, San Diego, CA, USA; DNA Clean & Concentrator kit, Zymo Research, Irvine, CA, USA) (see Appendix A, and Youssef O. et al., in preparation). To get the optimal amount of DNA (approximately 250 ng) for the downstream NGS, the restoration process was performed in duplicate per each FFPE tissue-derived DNA. The duplicates were pooled resulting in 50 μl of final pooled sample.

2.6. Assessment of the quality of FFPE tissue-derived DNA

DNA samples of the cohort 3 were processed for DNA extraction if not readily available. Aliquots taken from unrestored and restored FFPE tissue-derived DNA were sequenced for *GBA* variants by NGS. Sequencing results and performance metrics on per cent ≥20× coverage were used to monitor the impact of the restoration on the quality and performance of the FFPE tissue-derived DNA. Variants observed in the FFPE tissue-derived samples were compared to the variants determined from respective blood samples that were also sequenced, within the limits of sample availability, if mutation status was unknown.

Table 1
Electronic health record data sources and local adjustments to or exclusions of the original (Mehta et al., 2019; [16]) and additional (Mehta et al., 2020; [17]) signs and covariables utilised in the GED-C point scoring in the current study.

Original or modified GED-C sign/ covariable (adjustments to Mehta et al., 2019 [16] indicated)	Weighted scores	Data source (structural data/text mining/ both)	Laboratory and diagnosis code, and Finnish terms used in text mining [a]	GD patients [b, c]	Automated point scoring [b]
Splenomegaly, any extent (note the difference to the original scoring protocol)	3	Text	['suurentunut perna', 'pernan suurentuma', 'hypersplenismi', 'suuri perna', 'splenomegalia', 'kookas perna', 'laajentunut perna', 'suurentuneen pernan']	X	X
Disturbed oculomotor function (slow horizontal saccades with unimpaired vision)	3	Text	['silmän liikkeiden häiriö', 'silmävärve', 'nystagmus', 'silmän liikehäiriö', 'nykäisyliike']	X	X
Thrombocytopenia, mild or moderate (platelet count, 50–150 × 10^9/L)	2	Mainly structural (additional data text mined)	B -Trom 50–150 × 10^9/L (['trombosytopenia', 'trombopenia'])	X	X
Bone issues, including pain, crises, avascular necrosis, and fractures	2	Text	['luukipu', 'luukriisi', 'luusto%kuolio', 'luukipu']	X	X
Family history of Gaucher disease	2	Text	Manual chart review	X	–
Anaemia, mild or moderate (haemoglobin, 95–140 g/L)	2	Structural	B -Hb 95–140 g/L	X	X
Hyperferritinaemia, mild or moderate (serum ferritin, 300–1,000 μg/L)	2	Mainly structural (plasma ferritin; additional data text mined)	P -Ferrit 300–1,000 μg/L (['rautalasti', 'rautakuorma', 'hyperferrit%nemia'])	X	X
Jewish ancestry	2	NA	–	–	–
Disturbed motor function (impairment of primary motor development)	2	Text	['parkinsonismi', 'vapina', 'jäykkyys', 'spasmit']	X	X
Hepatomegaly, any extent (note the difference to the original scoring protocol)	2	Text	['suurentunut maksa', 'maksan laajentuma', 'hepatosplenomegalia', 'suuri maksa', 'maksa "melko kookas"']	X	X
Myoclonus epilepsy	2	Text	['myoklonaalinen%epilepsia', 'myoklooninen%epilepsia', 'epilepsia%myoklonaalinen', 'epilepsia%myokloninen']	X	X
Kyphosis	2	Text	['kyfoosi', 'selkäkyttyrä', 'kyfoottinen ryhti', 'th-rangan nikam%spontaani murtuma', 'th-rangan nikam%spontaani luhistuminen', 'äkkijyrkkä mutka rangassa']	X	X
Adult gammopathy – monoclonal or polyclonal	2	Both	ICD-10 D47.2, D89.0 ['Immunoglobuliinien pitoisuuden häiriö', 'hypergammaglobulinemia', 'gammapatia']	X	X
Anaemia, severe (haemoglobin, < 95 g/L)	1	Structural	B -Hb <95 g/L	X	X
Hyperferritinaemia, severe (serum ferritin, > 1,000 μg/L)	1	Mainly structural (additional data text mined)	P -Ferrit >1,000 μg/L (['rautalasti', 'rautakuorma', 'hyperferrit%nemia'])	X	X
Thrombocytopenia, severe (platelet count, < 50 × 10^9/L)	1	Structural	B -Trom <50 × 10^9/L.	X	X
Gallstones	0.5	Text	['sappikivet', 'sappikiviä']	X	X
Bleeding, bruising or coagulopathy	0.5	Both	ICD-10 D68 ['vuototaipumus']	X	X
Leukopenia	0.5	Mainly structural (additional data text mined)	B -Leuk <5 × 10^9/L (['leukopenia'])	X	X
Cognitive deficit	0.5	Both	ICD-10 R41.8 ['kognitiohäiriö', 'oppimishäiriö', 'kognitiivinen häiriö']	X	X
Low bone mineral density	0.5	Both	ICD-10 M80-M82 ['osteopenia', 'osteoporoosi', 'osteolyysi', 'luubiopsiassa nekroosia', 'hankalat luustomuutokset', 'luustonekroosi']	X	X
Growth retardation including low body weight	0.5	Text	['kasvuhäiriö', 'pienipainoisuus', 'alipainoisuus', 'lyhytkasvuisuus']	X	X
Asthenia	0.5	Text	['astenia']	X	X
Cardiovascular calcification	0.5	Both	ICD-10 I70 ['ateroskleroosi', 'sydämen ja verisuonten kalkkeutuminen', 'kalkkeutuminen hiippaläpän takapurjeessa', 'Hypertensio arterialis']	X	X
Dyslipidaemia	0.5	Both	ICD-10 code E78 ['dyslipidemia', 'hyperkolesterolemia', 'hypertriglyseridemia', 'rasva-aineenvaihdunnan häiriö']	X	X
Elevated angiotensin-converting enzyme levels	0.5	Structural	fS-ACE >65 U/L	X	X
Fatigue	0.5	Both	ICD-10 code R53 ['väsymys', 'lihasheikkous', 'uupumus', 'fatiikki']	X	X
Pulmonary infiltrates	0.5	Text	['keuhkojen varjostumat', 'keuhkoinfiltraatti']	X	X
Age ≤ 18 years [at diagnosis]	0.5	Structural	≤18 vuotta	X	NA
Family history of Parkinson's disease	0.5	NA	–	–	–
Blood relative who died of foetal hydrops and/or with diagnosis of neonatal sepsis of uncertain aetiology	0.5	NA	–	–	–
Additional signs/covariables introduced in Mehta et al., 2020 [17]	**Weighted scores**	**Data source (structural data/text mining/ both)**	**Laboratory and diagnosis code, and Finnish terms used in text mining [a]**	**GD patients [b]**	**Automated point scoring [b]**
Lymph node enlargement	NA	Both		X (separately)	X (separately)

(*continued on next page*)

Table 1 (*continued*)

Original or modified GED-C sign/ covariable (adjustments to Mehta et al., 2019 [16] indicated)	Weighted scores	Data source (structural data/text mining/ both)	Laboratory and diagnosis code, and Finnish terms used in text mining [a]	GD patients [b, c]	Automated point scoring [b]
Splenectomy	NA	Text	ICD-10 code R59 ['suurentuneet imusolmukkeet', 'imusolmukkeet suurentuneet', 'lymfadenopatia', 'lymphadenopatia'] ['splenektomia', 'poistettu perna', 'pernan poisto', 'perna poistettu']	X (separately)	X (separately)

Abbreviations: ACE, angiotensin converting enzyme; GED-C, Gaucher Earlier Diagnosis Consensus; Hb, haemoglobin; ICD-10, International Statistical Classification of Diseases and Related Health Problems, 10th Revision; NA, not applicable.

[a] "%" refers to any letters between the two parts of the phrase.

[b] "X" indicates that the sign/covariable was included in the scoring of the cohort(s) in question.

[c] Point-scored data was restricted to pretreatment time of GD patients.

Table 2
List of the *GBA* variants analysed in the current study.

Chr[1]	Position[1]	Reference SNP cluster ID	Variant	Clinical significance in ClinVar	Allele frequency (gnomAD[2])	Finnish allele frequency (based on 20,000–25,000 samples; gnomAD[2])	Further information	Availability in each chip version[3]
1	155235002	rs75822236	c.1604G > A	Pathogenic	0.0001694	0.000		v1
1	155235195	rs80356772	c.1505G > A	Pathogenic/ Likely pathogenic	0.000007963	0.000		v1
1	155235196	rs80356771	c.1504C > T	Pathogenic	0.00006724	0.000		v1/v2
1	155235205	rs369068553	c.1495G > C	Likely pathogenic	0.00005176	0.000		v1/v2
1	155235252	rs421016	c.1448 T > C	Pathogenic	0.001226	0.001837	1) Genotype analysed from cluster plots validated by NGS. 2) Hot spot variant; most common among Caucasians.	v1/v2
1	155235726	rs77369218	c.1343A > T	Likely pathogenic	–	–		v1
1	155235772	rs80356769	c.1297G > T	Pathogenic/ Likely pathogenic	0.00003182	0.000		v1
1	155235798	rs772548282	c.1271 T > C	Likely pathogenic	–	–		v1/v2
1	155235843	rs76763715	c.1226A > G	Pathogenic	0.002235	0.001314	1) Genotype analysed from cluster plots validated by NGS. 2) Hot spot variant; most common among Ashkenazi.	v1/v2
1	155236367	rs374306700	c.1102C > T	Likely pathogenic/ VUS	0.00001193	0.000		v1/v2
1	155237427	rs770796008	c.913C > G	Likely pathogenic	0.000003981	0.000		v1/v2
1	155237438	rs140955685	c.902G > A	VUS	0.0001097	0.000		v1/v2
1	155238228	rs61748906	c.667 T > C	Pathogenic/ Likely pathogenic	0.000007965	0.000		v1
1	155238597	rs398123530	c.508C > T	Pathogenic	0.000003981	0.00004619		v1
1	155240660	rs387906315	c.84dupG	Pathogenic	0.00004958	0.000		v1/v2

Abbreviations: Chr, chromosome; NGS, next-generation sequencing; SNP, small nucleotide polymorphism; VUS, a variant of uncertain significance.

[1] GRCh38.

[2] Allele frequencies of the variants according to gnomAD v2. 1.1 [20].

[3] Chip v1 and v2 have been utilised in the analysis of the samples in the Data freeze 4 and 5, respectively.

2.7. Sequencing

Sequencing was performed at Blueprint Genetics (Espoo, Finland). DNA samples were sequenced for *GBA* variants using NGS followed by Sanger sequencing, where needed, and according to the pipelines of Blueprint Genetics. NM_000157.4 was used as a reference transcript. All information regarding potential or confirmed pathogenic variants was provided to the geneticist and the clinical expert of Blueprint Genetics to include the variants of interest in the result statement regardless of the quality of the analysis.

DNA sequence data was not submitted to GenBank. Privacy policy and the material transfer agreement do not allow the transfer of the data to other registries or outside of EU and ETA. Furthermore, data on *GBA* variants only was allowed to be collected and published in the context of the current study.

2.8. Role of the funding source

The representative 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 Act. Only the personnel of the participating biobanks and Blueprint Genetics (only pseudonymised *GBA* sequencing data) had full access to the patient data. Corresponding author had final responsibility for the decision to submit for publication.

3. Results

3.1. GED-C point scoring of Helsinki Biobank sample donors previously diagnosed with GD

In Finland, the prototype GED-C PSS for GD has been evaluated in five Finnish GD patients diagnosed in OUH and enrolled to the previous study [15]. Respective EHR data allowed the point scoring before potential treatment. In the current study, additional GD patients, identified from Helsinki Biobank (HBB) sample donors, were point scored to obtain more data on the GED-C score range possibly indicating GD in the automatic point scoring of Finnish populations. Two cohorts were formed in HBB. Cohort 1 included patients previously diagnosed with GD, and cohort 2 consisted of patients examined for features of GD ("suspected GD patients") in HUH (Fig. 1). Altogether three previously diagnosed GD type 1 or 3 patients identified in HBB were eligible to the study (the cohort 1), all representing prospective biobank sample donors (a consent provided, and DNA extracted from blood and biobanked). Based on available EHR data, two of the three included GD patients also had a genetically confirmed diagnosis with respective data also being available (one homozygous for c.1448 T > C and the other compound heterozygote for c.1226A > G and c.681 T > G). In the current study, pathogenic *GBA* variants were confirmed by *GBA* NGS of blood DNA for the remaining prospective sample donor who had a clinical GD diagnosis but no previous genetic results available (a compound heterozygote for c.1448 T > C and c.1226A > G).

In addition, one "suspected GD patient" (the cohort 2) was eligible to the study. The patient had histological specimens previously analysed for morphological features associated with GD, but blood DNA was not available. Therefore, potential *GBA* variants(s) were analysed from an FFPE tissue specimen (see 3.2. Evaluation of the quality of FFPE tissue-derived DNA in *GBA* NGS). *GBA* NGS revealed pathogenic *GBA* variants, c.1448 T > C and c.1226A > G, thus confirming a genetic diagnosis for this patient.

Point scoring of the cohorts 1 and 2 was carried out in an automated manner, but also verified manually, where needed (Table 1; GD patients). Available EHR data of the secondary/tertiary health care of HUS allowed the point scoring with altogether 28 of the original 32 GED-C signs/covariables (Table 1; GD patients). EHR data available in HUS for the "suspected GD patient" (here genetically confirmed as a GD patient) allowed the assessment of only one GED-C laboratory variable (anaemia; mild or moderate); thus, the data of the patient was incomplete for consistent point scoring. The overall point-score range among the three GD patients of the cohort 1 was 12.5–22.5. Therefore, the patients in the current study had relatively higher point scores than the patients analysed in the previous study (point-score range 6–18.5). Table 3 shows the most prevalent signs/covariables observed in more than one previously diagnosed GD patients. Of note, all the previously diagnosed GD patients in the current study had splenomegaly, which is one the two signs weighted with a maximum of three points in the GED-C PSS. Furthermore, all patients had thrombocytopenia (mild or moderate), hyperferritinaemia (mild or moderate), low bone mineral density, and elevated angiotensin-converting enzyme level (Table 3). Regarding two additional signs introduced in Mehta et al. (17), i.e., splenectomy and lymph node enlargement, only one previously diagnosed GD patient had undergone splenectomy.

Table 3
The most prevalent GED-C signs/covariables observed among the previously diagnosed GD patients in the current study ($N = 3$).

GED-C sign/covariable	Points	N (max) = 3
Splenomegaly	3	3
Thrombocytopenia mild or moderate	2	3
Hyperferritinaemia mild or moderate	2	3
Low bone mineral density	0.5	3
Elevated angiotensin-converting enzyme levels	0.5	3
Anaemia mild or moderate	2	2
Family history of GD	2	2
Hepatomegaly	2	2
Bleeding, bruising or coagulopathy	1	2
Dyslipidaemia	0.5	2

3.2. Evaluation of the quality of FFPE tissue-derived DNA in GBA NGS

The suitability of samples can be critical for the screening approach where retrospectively identified potential undiagnosed GD patients are validated using biological specimens readily available in the biobanks. The performance of DNA extracted from potentially long-term stored FFPE tissue specimens in *GBA* NGS has previously not been studied in this context. In the current study, FFPE tissues of the cohort 3, namely the previously diagnosed/suspected GD patients (Fig. 1), were processed for DNA extraction followed by DNA restoration to improve the quality and, thus, the performance of the DNA in a subsequent NGS. One FFPE specimen was extracted per subject due to limited number of samples. Altogether six FFPE DNA specimens met the threshold for successful extraction.

The impact of the restoration on the quality and performance of FFPE samples in NGS was evaluated by utilising quality metrics data of NGS as well as sequencing results. Variants observed in the FFPE tissue-derived DNA samples were compared to the variants determined from respective high-quality blood DNA that were also sequenced if data on previous genetic tests wasn't readily available (Table 4). One of the analysed FFPE samples represented the "suspected GD patient" for whom neither previous data on *GBA* tests nor blood DNA for testing were available (HBP8; Table 4). Furthermore, two of the restored samples were contaminated before sequencing, and it was not possible to repeat the extraction and restoration for these samples. Therefore, the final analysis included altogether six unrestored and four restored samples (Table 4).

Based on the per cent ≥20× coverage in the NGS, restoration of the FFPE tissue-derived DNA improved the quality of the *GBA* NGS (Table 4). The analysis of the restored DNA of the "suspected GD patient" revealed pathogenic *GBA* variants, c.1448 T > C and c.1226A > G, thus confirming the genetic diagnosis for this patient. Overall, pathogenic variants, either previously known or confirmed in the current study, were found in five out of six unrestored and in all four restored FFPE tissue-derived DNA samples (Table 4). These data suggest that even long-term stored FFPE tissue specimens hosted by the Finnish biobanks represent a promising sample source for retrospective *GBA* NGS.

3.3. GED-C point scoring and SNP chip genotype analysis of the Helsinki Biobank sample donors

The preliminary GED-C point-score range in Finnish GD patients and a high number of Auria biobank sample donors whose GED-C point scores were in this range in the previous study highlighted a need for tools that could be used in further prioritisation of high-score biobank subjects for diagnostics [15]. In the current study, HBB sample donors were screened by utilising both the automated GED-C PSS and the SNP chip genotype data from an ongoing FinnGen study [18]. Two approaches were utilised; point scoring of genotype-based subgroups and the SNP chip genotype analysis of high-score subjects (Fig. 2). Due to release schedules of FinnGen data, both Data freeze 4 (DF4) and Data freeze five (DF5) SNP chip genotype data sets were utilised covering up

Table 4
The performance of DNA extracted from formalin-fixed paraffin-embedded tissues with and without DNA restoration in the next-generation sequencing of *GBA* gene (final $N = 6$).

ID	Gaucher status	Previously determined *GBA* variants	*GBA* variant status determined from blood DNA in the current study	FFPE tissue availability	FFPE age (years)	Successful of the *GBA* NGS			
						Unrestored FFPE DNA		Restored FFPE DNA	
						GBA variants	Per cent ≥ 20× coverage	*GBA* variants	Per cent ≥ 20× coverage
HBB1	Patient (the cohort 1)	c.1448T>C ho	–	Skin	10–15	c.1448T>C ho	63.12%	c.1448T>C ho	100.00%
HBB3	Patient (the cohort 1)	–	c.1448T>C he c.1226A>G he	Muscle	5–10	c.1448T>C he c.1226A>G he	98.96%	NA (Contamination in sampling)	
HBB8	Suspected (the cohort 2)	–	NA (DNA not available)	Lung	15–20	Unsuccessful	73.71%	c.1448T>C he c.1226A >G he	100.00%
BB1	Patient	c. 1226A>G ho	–	Thyroid	20–25	c. 1226A>G ho	94.04%	c. 1226A>G ho	100.00%
BB2	Patient	c.1448T>C he c.1226A>G he	–	Endometrium	20–25	c.1448T>C he c.1226A>G he	100.00%	c.1448T>C he c.1226A>G he	100.00%
BB4	Patient	c. 1226A>G ho	–	Appendix	10–15	c. 1226A>G ho	100.00%	NA (Contamination in sampling)	

Abbreviations: BB, Biobank Borealis; FFPE, formalin-fixed paraffin-embedded; GD, Gaucher disease; HBB, Helsinki Biobank; he, heterozygote; ho, homozygote; NA, not applicable; NGS, next-generation sequencing.

to ca 34,200 and 47,600 HBB sample donors, respectively, from which approximately 45,100 in overall had EHR data in the records of the secondary/tertiary health care of HUS (the cohort 4). Because SNP chip calling algorithms are generally not reliable for genotyping very rare variants [21], *GBA* genotypes were visually determined from genotyping cluster plots, and, where separately indicated, a subset of genotype data was assessed (see 2.4. Analysis of *GBA* genotypes from the SNP chip genotype data generated in the FinnGen study). The cluster plot analysis included all pathogenic or likely pathogenic *GBA* variants that were possible to analyse from the DF4 and DF5 data sets (Table 2). The genotypes that were considered potential homozygotes/compound heterozygotes based on the cluster plot analysis, if existed, were validated by *GBA* NGS throughout the study.

The SNP chip genotype data was available for all three previously diagnosed GD patients of the cohort 1. Additional potential homozygotes/compound heterozygotes for pathogenic/likely pathogenic *GBA* variants (potential undiagnosed GD patients) as well as potential heterozygote carriers of the most frequent pathogenic hot spot variants, c.1448 T > C and c.1226A > G, and respective negative controls were screened from a subset of HBB samples in the FinnGen DF4 (N ≈ 23,700/34,200) followed by the point scoring with respect to the three previously diagnosed GD patients (the cohort 1) and all sample donors in the FinnGen DF5 with the EHR data in HUS (N ≈ 45,100; the cohort 4) (Fig. 2; left pipeline).

The *GBA* variant statuses of all three previously diagnosed GD patients of the cohort 1 were predictable by the manual *GBA* cluster plot analysis of the SNP chip genotype data in terms of the variants included on the chip (Fig. 3). One GD patient (Fig. 3B) harboured a c.681 T > G variant, which is not covered by the FinnGen array chip and, thus, not possible to analyse. Additional homozygotes/compound heterozygotes for pathogenic/likely pathogenic *GBA* variants were not identified in the screened samples. There were only two additional samples initially suspected for compound heterozygosity, yet the subsequent sequencing confirmed negative *GBA* status for these two samples. Therefore, based on the analysis of a subset of samples in DF4 and variants possible to analyse on the FinnGen chip, altogether two out of four samples identified as potential homozygous/compound heterozygotes in the cluster plot analysis represented true carriers.

As expected, there were more samples, altogether 215 (0.9%) in the analysed 23,700 samples that were potentially heterozygous for a single pathogenic/likely pathogenic *GBA* variant based on the cluster plot analysis. From the 215 samples, 57 samples of 55 unique subjects represented potential carriers of one of the hot spot variants c.1448 T > C or c.1226A > G. To evaluate how many of the potential heterozygous carriers represented true heterozygotes or perhaps carry additional pathogenic/likely pathogenic *GBA* variants not captured by the FinnGen array, a subset of samples suspected for c.1448 T > C ($n = 8/34$) or c.1226A > G ($n = 4/23$) were sequenced by NGS. Sequencing confirmed that eight out of 12 sequenced samples were true heterozygote carriers of c.1448 T > C or c.1226A > G, and no additional pathogenic/likely pathogenic *GBA* variants were detected. In terms of all samples sequenced in the current study, it was estimated that the cluster plot analysis could predict c.1448 T > C and c.1226A > G carrier status with 62% and 100% accuracy, respectively. Based on these accuracies, the allele frequencies of c.1448 T > C and c.1226A > G variants in the FinnGen DF4 data set would be 0.001176 and 0.001196, respectively, which is well in line with the data in gnomAD (0.001837 and 0.001314, respectively; Table 2) [20].

Fig. 4 shows the relative distribution of the automated GED-C point scores (Table 1; Automated point scoring) among all the HBB samples in the FinnGen DF5 who also had EHR data available in HUS (N ≈ 45,100) and in each genotype-based subgroup that was possible to form based on the analysis a subset of the samples in the FinnGen DF4 (N ≈ 23,700/34,200), i.e., previously diagnosed GD patients of the cohort 1, potential heterozygote carriers of pathogenic hot spot variants c.1448 T > C and c.1226A > G ($N = 55$), and controls likely negative for pathogenic hot spot variants c.1448 T > C and c.1226A > G ($N = 4{,}670$). The point-score distribution of the genotype-based subgroups demonstrated that the higher the point score, more likely the sample represents potential GD patient, i.e., a homozygote or a compound heterozygote for pathogenic *GBA* variant(s) (Fig. 4). In general, the point scores among the 45,100 HBB subjects were variable, 0–17.5 points, with the majority (63.77%) of the assessed subjects having 0–2.0 points (Fig. 4). These results are in line with respective data from Auria biobank population [15].

In addition to a subset of HBB samples in the FinnGen DF4 (N ≈ 23,700/34,200) that were screened for additional potential homozygotes/compound heterozygotes, all HBB samples in the FinnGen DF5

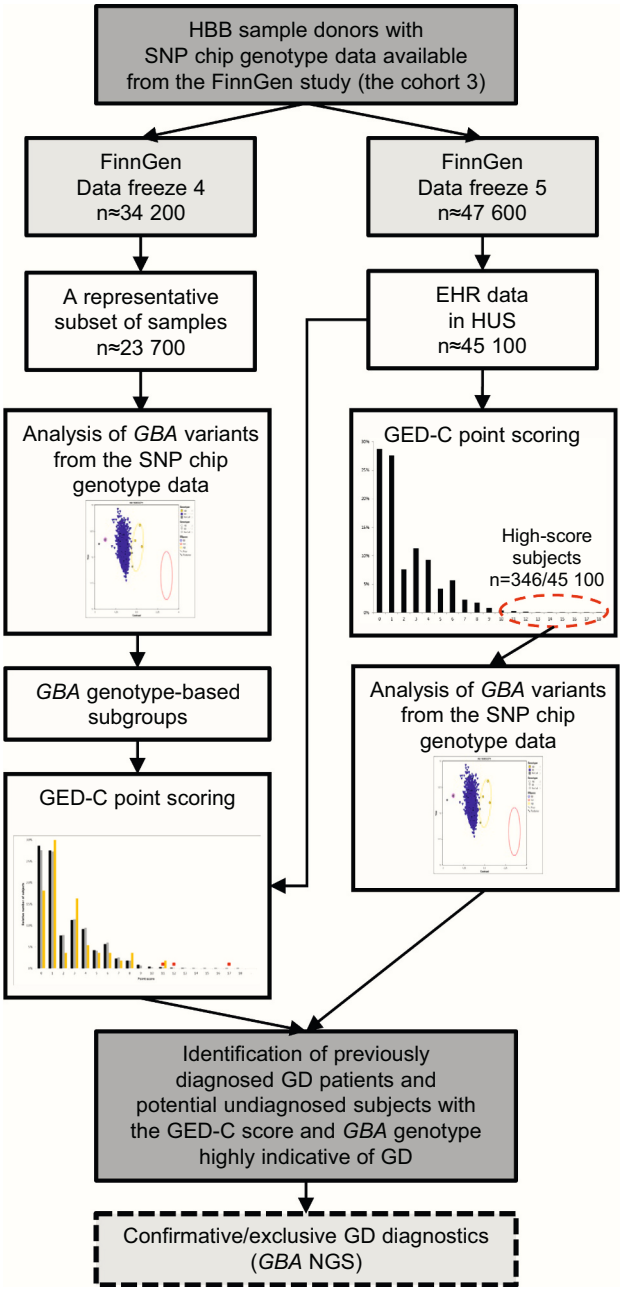

Fig. 2. A workflow of the screening for potential undiagnosed Gaucher disease (GD) patients in Helsinki Biobank (HBB). Both the automated GED-C point scoring and small nucleotide polymorphism (SNP) chip genotype data from the FinnGen study were utilised. Two genotype data sets were employed due to release schedules of the raw data (Data freeze 4 and 5).

with ≥10 points in the GED-C point scoring were also analysed for potential homozygotes/compound heterozygotes (n = 346/45,100; 0.77%) (Fig. 2; right pipeline). The cut-off was set to 10 based on the point scores of the previously diagnosed GD patients of the cohort 1. However, in addition to the previously diagnosed GD patients, none of the samples in the FinnGen DF5 with a score of ≥10 represented a potential homozygote/compound heterozygote for pathogenic/likely pathogenic *GBA* variants as determined based on the cluster plot analysis of the SNP chip genotype data. Instead, only four potential heterozygote carriers with a score of ≥10 were observed. Therefore, based on the GED-C point scoring and the analysis of the SNP chip genotype data of all high-score individuals, potential undiagnosed GD patients were not identified in the analysed 45,100 HBB sample donors.

Although potential undiagnosed GD patients were not identified in the screening of the assessed HBB subpopulation, these findings demonstrated that the SNP chip genotype data of the FinnGen study represents a valuable data source in further characterisation of the GED-C point-scored populations and allows prioritisation of samples for *GBA* diagnostics.

4. Discussion

The actual prevalence of GD in Finland is not known, and it has been hypothesized that GD is possibly underdiagnosed in this country. In accordance with worldwide attempts to identify potential undiagnosed patients [11–14], we have set up a retrospective screening approach in the unique Finnish biobank landscape that allows the screening of considerably large populations with data and samples readily available. The overall aim of the current and the previous study [15] was to utilise the GED-C PSS of GD type 1/3 [16] and data and samples available in Finnish biobanks in the screening of GD.

Finnish biobank sample collections, governed by the Finnish Biobank Act, can be utilised in rare disease screening. Data obtained in the current study suggest that even long-term stored FFPE samples represent a promising sample source for retrospective *GBA* diagnostic, thus increasing the applicability of biobank sample collections in the screening for e.g., undiagnosed GD patients. Further analyses with additional FFPE tissues samples are needed to evaluate on a broader scale the potential effects of tissue type and age of the sample.

Because a commonly accepted cut-off value of the GED-C PSS has not been assigned, the GED-C PSS was tested in previously diagnosed Finnish GD type 1/3 patients. Based on the data obtained from altogether eight patients of the previous and the current study, the Finnish GD patients may have high but variable GED-C point-score range, 6–22.5. The most often observed GED-C sign/covariable was mild or moderate anaemia, present in seven out of eight tested Finnish patients. Patients had other GED-C signs/covariables with varying frequencies. For example, splenomegaly, one of the two GED-C signs weighted with a maximum of three points, was recorded for all three patients in the current study, and for three out of five patients in the previous study [15]. It should be noted that the point scorings were not performed identically in the two studies due to differences in data availability, approaches, and terms used in the data extraction and text mining. Nevertheless, the data from eight previously diagnosed Finnish GD patients provides an indicative score range to be employed in automated screening for undiagnosed patients in Finland.

The majority of the GED-C PSS signs/covariables can be utilised in automated point scoring of Finnish secondary/tertiary health care data as demonstrated in the cohort of 160,000 sample donors of Auria Biobank (Turku, Finland) [15] and in Helsinki Biobank (Helsinki, Finland) subpopulation consisting 45,100 sample donors. The GED-C point score distributions in the two tested biobank populations/subpopulations were roughly in line and showed that although a majority of the screened subjects have low point scores, there is also a substantial number of individuals with median and high point scores. In fact, there were relatively more high-score patients in HBB subpopulation as the point-score range was wider, 0–17.5, compared to that observed in Auria biobank population, 0–13.5 points. This finding further emphasizes the role of additional characterisation of the high-score subjects in the screening of large populations. Indeed, the current study showed that the automated GED-C PSS alone is not sufficient for the most optimal identification of potential undiagnosed GD patients but can be a powerful tool when utilised together with a large genotype data set available from Finnish biobank sample donors. In the current study, the SNP chip genotype data on *GBA* variants, obtained from the FinnGen study, was utilised and analysed by visual cluster plot analysis to accompany GED-C point scoring results of the HBB sample donors. Previously diagnosed GD patients were correctly identified with this

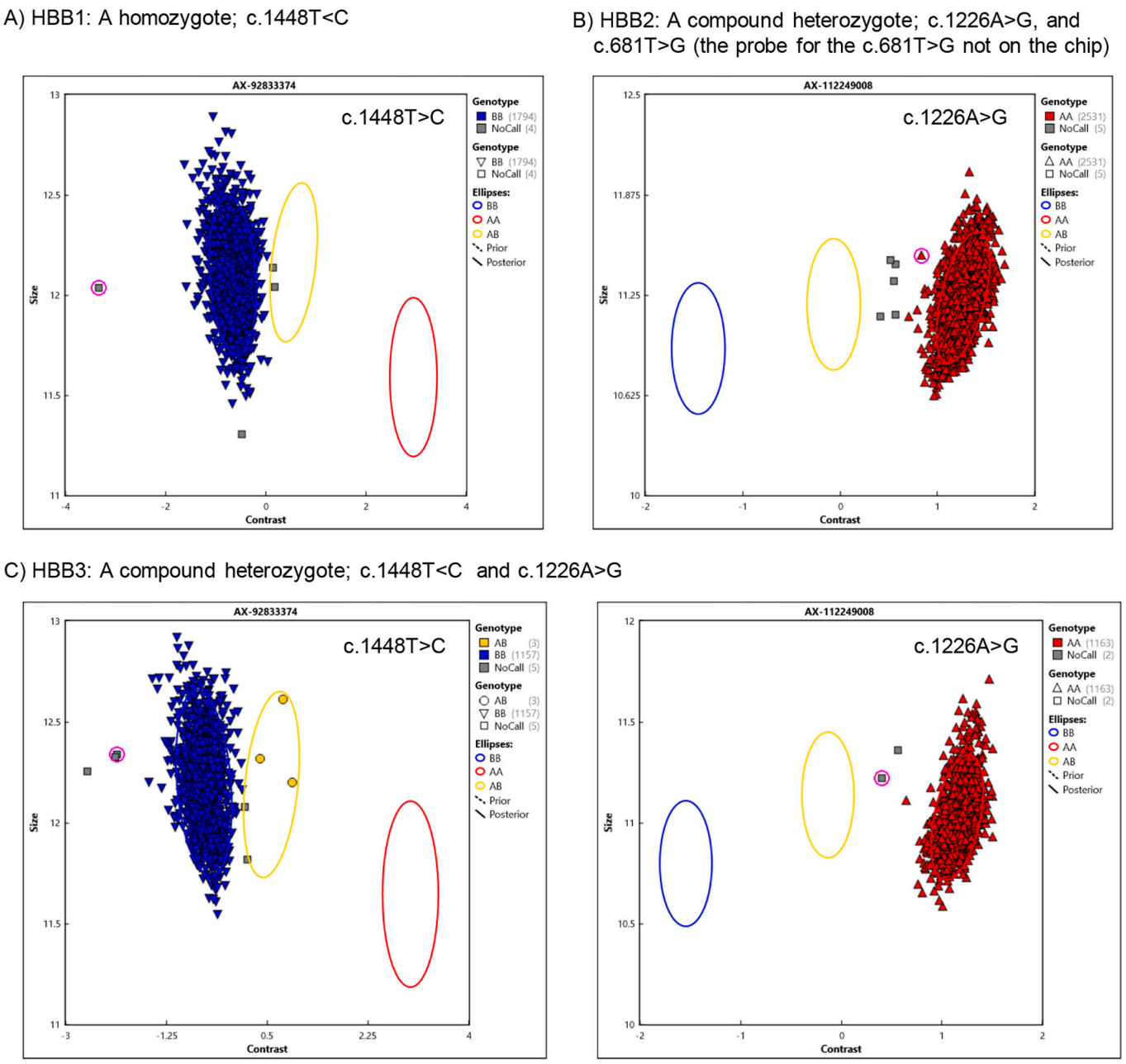

Fig. 3. Cluster plots generated from small nucleotide polymorphism (SNP) genotype data of Helsinki Biobank samples genotyped in the FinnGen study [18]. The plots shown here include the samples of the three previously diagnosed Gaucher disease patients identified in Helsinki Biobank and eligible to the study (HBB1–HBB3 in A–C, respectively; marked with pink circles). One plot corresponds to one variant, while one spot in each plot corresponds to one sample. The colour of the spots and ellipses indicate the computed genotype result and cluster boundaries, respectively, as determined by the SNP chip calling algorithm. The algorithm is not reliable for genotyping rare variants.

approach, but additional, undiagnosed GD patients, homozygotes or compound heterozygotes for pathogenic/likely pathogenic *GBA* variants, were neither identified in the analysed subset of HBB sample donors regardless of the GED-C score, nor detected among all individuals with the score of 10 or more. Data not included in the analyses of the SNP chip genotypes may contain potential undiagnosed patients with extremely low GED-C scores or undiagnosed patients with genotypes currently not covered by the FinnGen array chip. It should also be noted that, in terms of validation, the high-score individuals were neither sequenced for *GBA* mutations nor analysed for e.g., lyso-Gb1 levels independently from the cluster plot analysis of the SNP chip genotype data. However, the cluster plot analysis was validated for the most often recorded pathogenic/likely pathogenic variants observed in the Finnish population so far, and all potential homozygotes/compound heterozygotes, identified in the cluster plot analysis, were validated by *GBA* NGS. Therefore, it is unlikely that subjects with both the GED-C score and *GBA* genotypes highly indicative of GD were missed in the current study. It should be noted that a genetic diagnosis was yet confirmed for one suspected GD patient previously treated for GD-related features but who had incomplete data available for consistent point scoring.

The findings of the current and previous study imply that the number of undiagnosed GD patients is negligible in Finland, suggesting that the true prevalence of GD in Finland is close to the prevalence of diagnosed patients, ~1:325,000. Indeed, GD can be extremely rare in Finland due to unique genetic characteristics of the Finnish population, demonstrated by the variant allele frequencies ([20]; Table 2). However, the

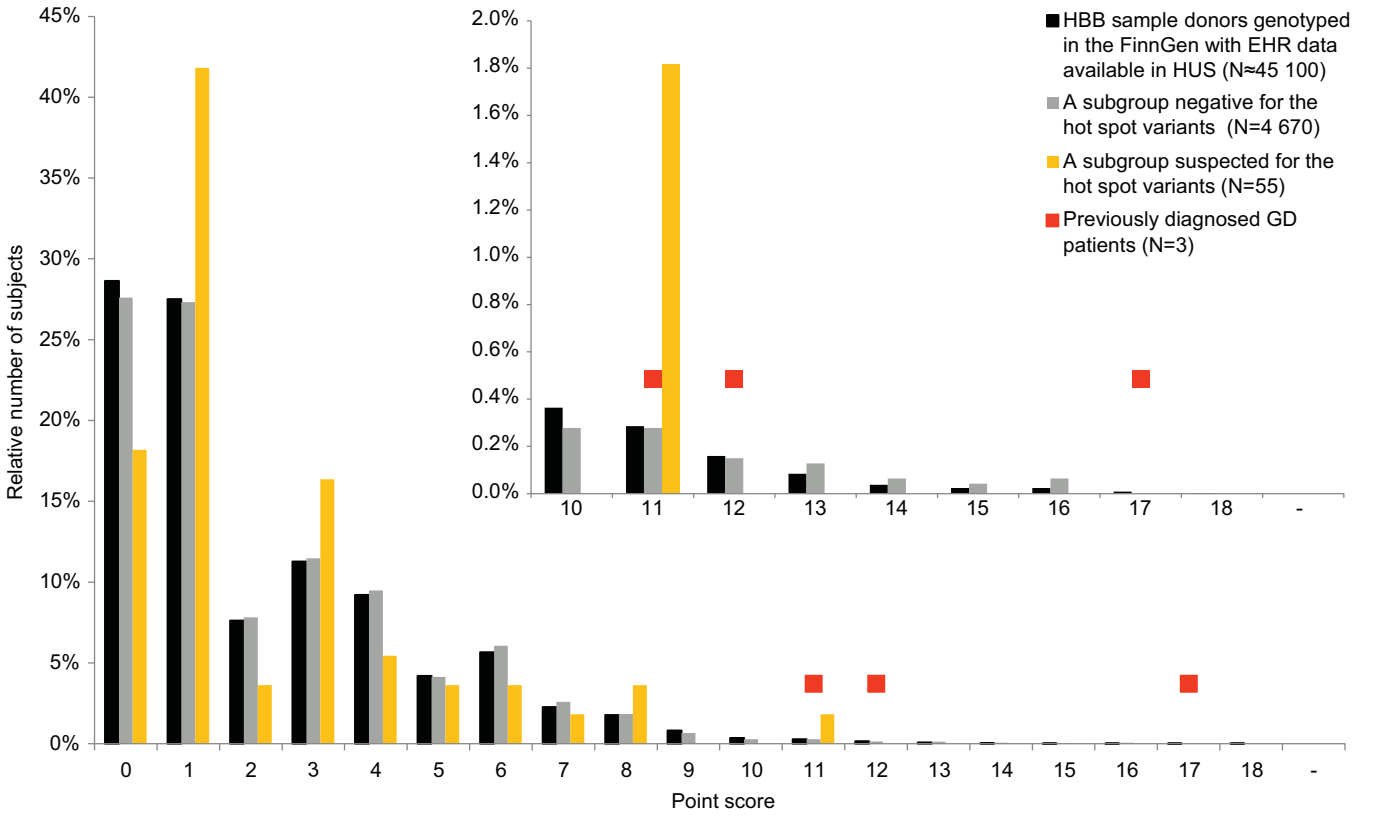

Fig. 4. The GED-C point-score distribution in Helsinki Biobank (HBB) subpopulations representing all HBB sample donors who have been genotyped in the FinnGen study and who have electronic health record data in the records of the Hospital District of Helsinki and Uusimaa (N ≈ 45,100; black columns), and in respective subgroups of individuals suspected ($N = 55$; yellow columns) or negative ($N = 4{,}670$; grey columns) for the *GBA* hot spot variants c.1448 T > C or c.1226A > G accompanied by the separately indicated point scores of the previously diagnosed GD patients who were identified in HBB and eligible to the study, and who have also been genotyped in the FinnGen ($N = 3$; red data points). The point scoring of all samples was carried out in an automated manner. The main image shows the distribution of values (rounded to closest integer) among all assessed subjects while the insert represents subjects with a score of ≥10 ($n = 346$).

occurrence of the carriers of pathogenic *GBA* variants in different parts of Finland can be variable, and, therefore, the prevalence and the number of potential undiagnosed patients, if existed, can differ between the different parts of the country [23]. Subjects in the HBB subpopulation, whose genotype was not suggestive for GD, but who presented with high GED-C scores, likely had other diagnoses/conditions (including potential combinations) resulting in signs/symptoms included in the GED-C PSS, including ICD-10 D69 Purpura and other haemorrhagic conditions, D50 iron deficiency anaemia, K80 Gallstones, M80 osteoporosis and pathological fracture, S32 lumbar vertebral or hip fracture, M87 Bone necrosis, and C90 Multiple Myeloma. Such patients were not characterised further in the current study.

This study is based on existing SNP chip genotype and EHR data and thereby limited by the availability of the data to be collected. The analysis of the pathogenic *GBA* variants was restricted to the variants covered by the FinnGen array chip. The secondary/tertiary health care data of remaining Finnish hospital districts and primary health care data in general were not utilised in the study. Furthermore, it is possible that in the data extraction process, information for all remaining GED-C PSS data fields was not equally available from all assessed subjects.

5. Conclusions

The current study demonstrated that the GED-C PSS, data and samples available in Finnish biobanks, and Finnish EHRs together represent an efficient way to screen for undiagnosed GD patients in Finland. The SNP chip genotype data can be used for the identification of potential carriers of GD-associated rare pathogenic variants among the GED-C point-scored subjects, but the analysis pipeline of the SNP chip genotype data should be developed for rare variants. The tools that were set up to screen for undiagnosed GD patients in Finnish biobank populations can be applied to the screening of other rare genetic diseases with known genetic background in the Finnish landscape.

Contributors

KE and OC contributed to study concept. All authors participated in study design and interpretation of the results. HH, UWK, and SM performed the GED-C point scoring with the assistance of other authors. MP carried out the genetic analysis of FinnGen genotype data. MP, PB, and KU coordinated the sample analytics together with the personnel of HBB and Blueprint Genetics. OC and the personnel of HBB and BB processed FFPE tissue specimens. Remaining sample processing was carried out by the personnel of HBB. HH, MP, and KU contributed to the design and drawing of figs. KU coordinated the preparation of the manuscript which was edited and approved by all authors.

Funding

This study was funded by Takeda.

Author statement

Term	Author	Definition
Conceptualization	KE, OC	

(*continued on next page*)

(*continued*)

Term	Author	Definition
		Ideas; formulation or evolution of overarching research goals and aims
Methodology	All	Development or design of methodology; creation of models
Validation	HH, UWK, and SM performed the GED-C point scoring. Aggregate data results (pseudonymised) were validated by all authors. MP carried out the genetic analysis of FinnGen genotype data that was partly confirmed by NGS.	Verification, whether as a part of the activity or separate, of the overall replication/reproducibility of results/experiments and other research outputs
Formal analysis	HH, MP, SM	Application of statistical, mathematical, computational, or other formal techniques to analyse or synthesize study data
Investigation	HH, MP, UWK, SM, PB, OC, the personnel of the participating biobanks	Conducting a research and investigation process, specifically performing the experiments, or data/evidence collection
Resources	Participating institutions/parties. Funding from Takeda.	Provision of study materials, reagents, materials, patients, laboratory samples, animals, instrumentation, computing resources, or other analysis tools
Data Curation	Participating biobanks and Medaffcon (published aggregate data).	Management activities to annotate (produce metadata), scrub data and maintain research data (including software code, where it is necessary for interpreting the data itself) for initial use and later reuse
Writing - Original Draft	KU	Preparation, creation and/or presentation of the published work, specifically writing the initial draft (including substantive translation)
Writing - Review & Editing	All	Preparation, creation and/or presentation of the published work by those from the original research group, specifically critical review, commentary or revision – including pre-or postpublication stages
Visualization	KU, HH, MP	Preparation, creation and/or presentation of the published work, specifically visualization/data presentation
Supervision	KE	Oversight and leadership responsibility for the research activity planning and execution, including mentorship external to the core team
Project administration	KE, KU, MIL	Management and coordination responsibility for the research activity planning and execution
Funding acquisition	Funding from Takeda	Acquisition of the financial support for the project leading to this publication

Declaration of Competing Interest

KE is employed by Takeda (Helsinki, Finland) and holds stocks/stock options in Takeda. HH, MP, SM, PB, and OC are employed by Helsinki Biobank (Helsinki, Finland) which received reimbursement from Medaffcon and Takeda, for the work done at the biobank. KU and ML are employed by Medaffcon Oy (Espoo, Finland) which received reimbursement from Takeda, for conducting the study. UWK reports consultancy fees from Medaffcon and Takeda during the study, as well as grant support, paid to her institution, from several foundations for research outside the submitted work. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability

The authors do not have permission to share data.

Acknowledgements

This study benefited from the samples/data from the Helsinki Biobank, Helsinki, Finland (https://www.helsinginbiopankki.fi) and the Biobank Borealis of Northern Finland, Oulu, Finland (https://www.ppshp.fi/Tutkimus-ja-opetus/Biopankki/Pages/default.aspx). Biobank sample donors and the participants and investigators of the FinnGen study are acknowledged. Dr. Omar Youssef is acknowledged for carrying out FFPE DNA restoration and sharing his technological expertise. The personnel of the biobanks and Blueprint Genetics are thanked for their valuable help. The study was funded by Takeda.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ymgmr.2022.100911.

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] A. Mehta, Epidemiology and natural history of Gaucher's disease, Europ. J. Intern. Med. 17 (2006) S2–S5, https://doi.org/10.1016/j.ejim.2006.07.005.

[5] P.J. Meikle, Prevalence of lysosomal storage disorders, JAMA. 281 (1999) 249, https://doi.org/10.1001/jama.281.3.249.

[6] 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, Orphan. J. Rare Dis. 7 (2012) 77, https://doi.org/10.1186/1750-1172-7-77.

[7] 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.

[8] P.K. Mistry, S. Sadan, R. Yang, J. Yee, M. Yang, Consequences of diagnostic delays in type 1 Gaucher disease: The need for greater awareness among Hematologists–Oncologists and an opportunity for early diagnosis and intervention, Am. J. Hematol. 82 (2007) 697–701, https://doi.org/10.1002/ajh.20908.

[9] 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.

[10] A.S. Thomas, A.B. Mehta, D.A. Hughes, Diagnosing Gaucher disease: an on-going need for increased awareness amongst haematologists, Blood Cell Mol. Dis. 50 (2013) 212–217, https://doi.org/10.1016/j.bcmd.2012.11.004.

[11] Splenomegaly Gaucher Group, I. Motta, D. Consonni, M. Stroppiano, C. Benedetto, E. Cassinerio, B. Tappino, P. Ranalli, L. Borin, L. Facchini, A. Patriarca, W. Barcellini, F. Lanza, M. Filocamo, M.D. Cappellini, Predicting the probability of Gaucher disease in subjects with splenomegaly and thrombocytopenia, Sci. Rep. 11 (2021) 2594, https://doi.org/10.1038/s41598-021-82296-z.

[12] I. Ntanasis-Stathopoulos, M. Gavriatopoulou, D. Fotiou, N. Kanellias, M. Migkou, E. Eleutherakis-Papaiakovou, E. Kastritis, M.-A. Dimopoulos, E. Terpos, Screening for Gaucher disease among patients with plasma cell dyscrasias, Leuk. Lymphoma 62 (2021) 761–763, https://doi.org/10.1080/10428194.2019.1672059.

[13] R.P. Limgala, V. Furtak, M.M. Ivanova, E. Changsila, F. Wilks, M.N. Fidelia-Lambert, O. Goker-Alpan, M.C. Gondré-Lewis, Selective screening for lysosomal storage disorders in a large cohort of minorities of African descent shows high prevalence rates and novel variants, JIMD Rep. 59 (2021) 60–68, https://doi.org/10.1002/jmd2.12201.

[14] T.M. Reynolds, A.S. Wierzbicki, V. Skrahina, C. Beetz, PATHFINDER project collaboration group, screening for patients with Gaucher's disease using routine pathology results: PATHFINDER (ferritin, alkaline phosphatase, platelets) study, Int. J. Clin. Pract. 75 (2021), https://doi.org/10.1111/ijcp.14422.

[15] M.J. Savolainen, A. Karlsson, S. Rohkimainen, I. Toppila, M.I. Lassenius, C. V. Falconi, K. Uusi-Rauva, K. Elomaa, 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, Mol. Genet. Metabol. Rep. 27 (2021), 100725, https://doi.org/10.1016/j.ymgmr.2021.100725.

[16] 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.

[17] 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.

[18] FinnGen, FinnGen. https://www.finngen.fi/en, 2022.

[19] FinnGen, FinnGen. Documentation. https://finngen.gitbook.io/documentation/, 2021.

[20] Consortium, The Genome Aggregation Database (gnomAD). https://gnomad.broadinstitute.org/, 2021.

[21] M. Weedon, L. Jackson, J. Harrison, K. Ruth, J. Tyrrell, A. Hattersley, C. Wright, Use of SNP chips to detect rare pathogenic variants: retrospective, population based diagnostic evaluation, BMJ. (2021), n214, https://doi.org/10.1136/bmj.n214.

[22] G.M. Pastores, D.A. Hughes, Gaucher disease, in: M.P. Adam, H.H. Ardinger, R. A. Pagon, S.E. Wallace, L.J. Bean, K.W. Gripp, G.M. Mirzaa, A. Amemiya (Eds.), GeneReviews®, University of Washington, Seattle, Seattle (WA), 1993. http://www.ncbi.nlm.nih.gov/books/NBK1269/ (accessed February 10, 2022).

[23] The Sequencing Initiative Suomi (SISu). http://www.sisuproject.fi/, 2016.