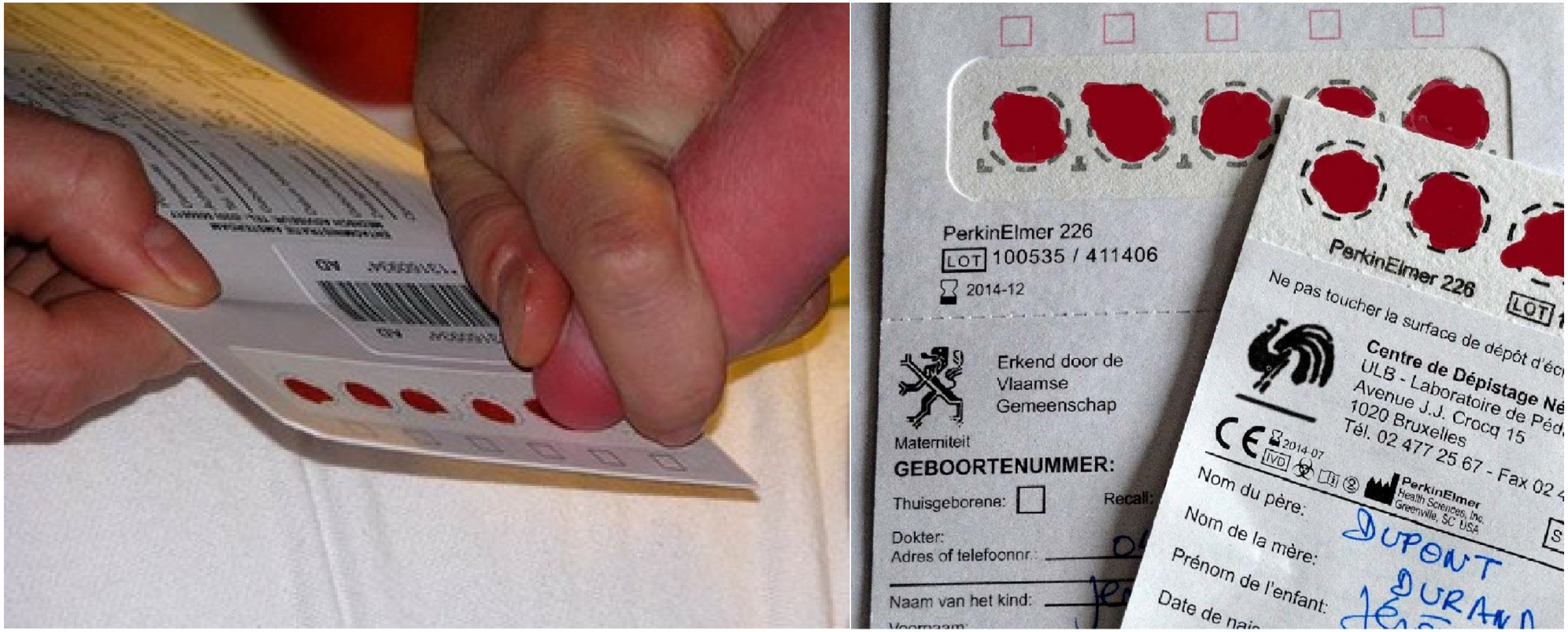
KCE REPORT 267

Federaal Kenniscentrum voor de Gezondheidszorg
Centre Fédéral d'Expertise des Soins de Santé
Belgian Health Care Knowledge Centre

MULTI CRITERIA DECISION ANALYSIS TO SELECT PRIORITY DISEASES FOR NEWBORN BLOOD SCREENING



2016 www.kce.fgov.be

KCE REPORT 267
HEALTH SERVICES RESEARCH

Federaal Kenniscentrum voor de Gezondheidszorg
Centre Fédéral d'Expertise des Soins de Santé
Belgian Health Care Knowledge Centre

MULTI CRITERIA DECISION ANALYSIS TO SELECT PRIORITY DISEASES FOR NEWBORN BLOOD SCREENING

CHRIS DE LAET, GERMAINE HANQUET, ERIK HENDRICKX

2016 www.kce.fgov.be

COLOPHON

Title: Multi Criteria Decision Analysis to select priority diseases for Newborn blood screening

Authors: Chris De Laet (KCE), Germaine Hanquet (KCE), Erik Hendrickx (KCE)

Project coordinator: Marijke Eyssen (KCE)

Reviewers: Irina Cleemput (KCE), Frank Hulstaert (KCE), Raf Mertens (KCE), Genevieve Veereman (KCE)

External experts: François Boemer (Université de Liège), Marc Boënne (BOKS, Belgische Organisatie voor Kinderen en volwassenen met een Stofwisselingsziekte / Organisation Belge pour des enfants et des adultes souffrant d'une maladie métabolique), Vincent Bours (Université de Liège), Karen Colaert (Vlaams Agentschap Zorg en Gezondheid), Lutgarde De Baere (BOKS), Linda De Meirleir (AZ-VUB), François Eyskens (Provinciaal Centrum voor de opsporing van metabole aandoeningen), Philippe Goyens (CHU Brugmann and HUDERF), Hilde Laeremans (AZ-VUB and CHU Brugmann), Marie-Christine Mauroy (Office de la Naissance et de l'Enfance or ONE), Heidi Mertes (Ugent), Marie-Cécile Nassogne (Cliniques universitaires St Luc), Tatiana Pereira (Fédération Wallonie Bruxelles (FWB) and ONE), Roland Schoos (Université de Liège), Elfriede Swinnen (Orphanet), Béatrice Toussaint (SPF Santé Publique / FOD Volksgezondheid), Pieter Vandenbulcke (Vlaams Agentschap Zorg en Gezondheid), Marie-Françoise Vincent (Cliniques universitaires St Luc), Véronique Zinnen (FWB and ONE).

External validators: Jean-Jacques Cassiman (KU Leuven), Corinne De Laet (HUDERF, Brussels), Brigitte Côté (Institut National d'Excellence en Santé et en Services Sociaux (Québec, Canada))

Acknowledgements: Muriel Deguerry (Brussels-Capital Health and Social Observatory) for providing information on NBS organisation in the Brussel Capital region. Stephan Devriese (KCE) for help with specific graphs. Gudrun Briat (KCE) and Karin Rondia (KCE) for review and adaptation of syntheses in Dutch and French. Cover pictures: De Kern – verloskundigenpraktijk (NL) and Andrée Mangin (KCE).

Other reported interests: None declared: François Boemer (Université de Liège), Jean-Jacques Cassiman (KU Leuven), Corinne De Laet (HUDERF, Brussels), Linda De Meirleir (AZ-VUB), François Eyskens (Provinciaal Centrum voor de opsporing van metabole aandoeningen), Heidi Mertes (UGent), Marie-Françoise Vincent (Cliniques universitaires St Luc) and Véronique Zinnen (FWB and ONE).

Membership of a stakeholder group on which the results of this report could have an impact: Hilde Laeremans (ULB en Vlaams Centrum Brussel voor opsporing van aangeboren Metabole Aandoeningen), Tatiana Pereira (FWB, Direction Générale de la Santé Exécutant pour ONE) and Roland Schoos (Université de Liège, Centre de dépistage néonatal de Liège).

Presidency or accountable function within an institution, association, department or other entity on which the results of this report could have an impact: Karen Colaert (organising the Flemish screening programme "aangeboren aandoeningen bij pasgeborenen via een bloedstaal" for the Vlaams Agentschap Zorg en Gezondheid) Lutgarde De Baere (president of the BOKS patient association for metabolic diseases), Philippe Goyens and Hilde

Laeremans (coordinating a laboratory for screening of metabolic disorders for the ULB and the AZ-VUB (VCBMA), Roland Schoos (participating to the Steering Committee for neonatal screening of metabolic disorders of the FWB), Béatrice Toussaint (responsible of neonatal screening in the cabinet of the FWB in 2009-2014), Pieter Vandenbulcke (head of the team general prevention at the Vlaams Agentschap Zorg en Gezondheid and responsible for population screening programmes related to disease prevention).

Other possible interests that could lead to a potential or actual conflict of interest:

Brigitte Côté (participated as principal investigator to the INESSS scientific study (2013) "Pertinence d'élargir le programme de dépistage néonatal sanguin au Québec")

Layout:	Joyce Grijseels, Sophie Vaes.
Disclaimer:	• **The external experts were consulted about a (preliminary) version of the scientific report. Their comments were discussed during meetings. They did not co-author the scientific report and did not necessarily agree with its content.** • **Subsequently, a (final) version was submitted to the validators. The validation of the report results from a consensus or a voting process between the validators. The validators did not co-author the scientific report and did not necessarily all three agree with its content.** • **Finally, this report has been approved by common assent by the Executive Board.** • **Only the KCE is responsible for errors or omissions that could persist. The policy recommendations are also under the full responsibility of the KCE.**
Publication date:	25 april 2016
Domain:	Health Services Research (HSR)
MeSH:	Decision Making; Metabolism, Inborn Errors; Neonatal Screening
NLM Classification:	WD 205
Language:	English
Format:	Adobe® PDF™ (A4)
Legal depot:	D/2016/10.273/44
ISSN:	2466-6459
Copyright:	KCE reports are published under a "by/nc/nd" Creative Commons Licence http://kce.fgov.be/content/about-copyrights-for-kce-reports.

How to refer to this document?	De Laet C, Hanquet G, Hendrickx E. Multi Criteria Decision Analysis to select priority diseases for Newborn blood screening. Health Services Research (HSR) Brussels: Belgian Health Care Knowledge Centre (KCE). 2016. KCE Reports 267. D/2016/10.273/44. This document is available on the website of the Belgian Health Care Knowledge Centre.

■ TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ABBREVIATION	DEFINITION
17-OHP	17-alpha-hydroxyprogesterone
ACADVL	Acyl-CoA dehydrogenase, very long chain.
BTD	Biotinidase
C14:2	Tetradecenoylcarnitine
CAH	Congenital Adrenal Hyperplasia
CBS	Cystathionine Beta Synthase
CHT	Congenital Hypothyroidia
CF	Cystic Fybrosis (mucoviscidose)
ETF	Electron-transferring-flavoprotein
FAH	FumarylAcetoacetase
FAO	Fatty Acid Oxidation
FWB	Fédération Wallonie-Bruxelles (Belgium: French speaking community)
FN	False Negative (for screening test)
FP	False Positive (for screening test)
GA-1	Glutaric aciduria type 1
GAL	Galactosaemia (all types)
GALE	UDP galactose epimerase
GALK	Galactokinase
GALT	Galactose-1-phosphate uridyl transferase
HCY	Homocystinuria
IV	Intra Venous
IVA	Isovaleric acidemia
INESSS	Institut National d’Excellence en Santé et en Services Sociaux (Québec, Canada)
INSPQ	Institut National de Santé Publique du Québec
LMCD	Late-onset, biotin-responsive, Multiple Carboxylase Deficiency (Biotinidase deficiency)
MADD	Multiple Acyl-CoA dehydrogenase deficiency (glutaric aciduria type II)

MCAD	Medium chain Acyl-CoA dehydrogenase deficiency
MCDA	Multi-Criteria Decision Analysis
MCT	Medium-Chain Triglycerides
MMA	MethylMalonic Acidemia
MS/MS	Tandem mass spectrometry
MSUD	Maple Syrup Urine Disease (Leucinosis)
NBS	Newborn Blood Screening
NPV	Negative Predictive Value (for test)
NSC	National Screening Committee (UK)
NTBC	Nitisinone (NTBC is an abbreviation of the full chemical name)
PA	Propionic Acidemia
PKU	Phenylketonuria
PPV	Positive Predictive Value (for test)
SUAC	Succinylacetone
TN	True Negative (for screening test)
TP	True Positive (for screening test)
TYR I	Tyrosinemia Type I (also called Hereditary Tyrosinemia type I, HTI)
AZG	Agentschap Zorg en Gezondheid (Belgium, Dutch speaking community
VG	Vlaamse Gemeenschap (Belgium: Dutch speaking community)
VLCAD	Very long-chain acyl-CoA dehydrogenase

■ SCIENTIFIC REPORT

1 INTRODUCTION AND SCOPE

The Newborn Blood Screening (NBS) programme is a public health programme intended to systematically screen all infants shortly after birth for a list of conditions that are treatable, but not clinically evident in the newborn period. Most diseases included in NBS are inborn metabolic diseases whose first symptoms appear during the neonatal period or early childhood, after irreversible damage has occurred. Screening for these diseases before symptoms appear usually allows an early diagnosis and early interventions. Neonatal screening is also commonly called the neonatal heel prick or Guthrie test. A simple small blood sample is taken and the blood is soaked into a pre-printed collection card known as Guthrie card (see Figure 1).

Internationally, there is no clear consensus on which inborn diseases need to be screened. As a result, the number and nature of included diseases varies enormously by country, from none to over forty diseases.

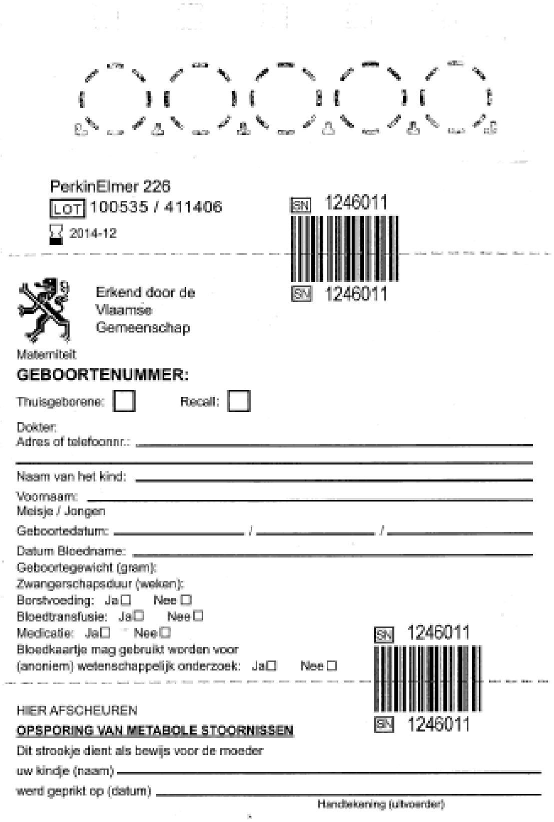
In Belgium, a national NBS programme started in 1968 with the systematic screening of all newborns for one metabolic disease (phenylketonuria). Six other diseases have been progressively added to the programme.[2, 3] In the early eighties, the responsibility of the programme has been transferred to the communities.

Most inborn metabolic diseases (but not all) are today detected by tandem mass spectrometry (MS/MS), a laboratory technique able to screen for multiple metabolic disorders simultaneously and rapidly, through the analysis of a single blood sample. This technique detects abnormal levels of a high number of metabolites and must be followed by confirmatory tests for final diagnosis. In the VG the total cost for the primary screening (including cards, mailing, reporting of results and education material) amounts to around 20 € per newborn.

As the MS/MS technique allows to detect a high number of disorders, programmes must decide which disease should be screened for, aiming at an acceptable balance between benefits and risks. Possible benefits of screening for a disease are that early detection followed by effective intervention can prevent illness, sequelae and in some diseases early death. The main risks are the impact of false negative and false positive results, involving false reassurance or unnecessary worry and costs respectively.

As a result of this technological improvement in the last years, some countries enlarged considerably the list of diseases to be screened. As a result, there are large variations in the diseases screened for in European countries and no validated rules for decision making decisions on how to expand NBS programmes,[1] This lead the Flemish community to ask KCE to conduct a study on this topic to explore methods to structure this debate.

Figure 1 – Sample Guthrie test card



PerkinElmer 226
LOT 100535 / 411406
2014-12

SN 1246011
SN 1246011

Erkend door de
Vlaamse
Gemeenschap

Materniteit
GEBOORTENUMMER:

Thuisgeborene: □ Recall: □

Dokter:
Adres of telefoonnr.: ____________________

Naam van het kind: ____________________
Voornaam: ____________________
Meisje / Jongen
Geboortedatum: ________ / ________ / ________
Datum Bloedname: ____________________
Geboortegewicht (gram):
Zwangerschapsduur (weken):
Borstvoeding: Ja□ Nee □
Bloedtransfusie: Ja□ Nee □
Medicatie: Ja□ Nee□
Bloedkaartje mag gebruikt worden voor
(anoniem) wetenschappelijk onderzoek: Ja□ Nee □

SN 1246011
SN 1246011

HIER AFSCHEUREN
<u>OPSPORING VAN METABOLE STOORNISSEN</u>
Dit strookje dient als bewijs voor de moeder
uw kindje (naam) ____________________
werd geprikt op (datum) ____________________
Handtekening (uitvoerder)

Since then, the list of diseases to be screened is decided by each community, according to their own legal criteria. As a result, the two communities do not screen for the same set of diseases.

The scope of this study is to pilot test a method to prioritize diseases to be included in the existing neonatal blood screening programs in both communities in Belgium. This is done through a process of Multi-Criteria Decision Analysis (MCDA) using a number of criteria that are weighed by a limited number of stakeholders (community decision makers, labs performing the tests, patient groups and ethical experts). Specific diseases under assessment are subsequently scored on each of these criteria and ranked.

Because this is a pilot test to evaluate the potential of this approach, it was limited to the six specific diseases that are screened in only one of the Belgian communities and not in the other. If successful, this methodology can be used to assess a larger set of diseases for the NBS programme.

Therefore, this report should not be interpreted as a final recommendation for in- or exclusion of a disease in the screening programme.

2 NEWBORN BLOOD SCREENING FOR INBORN ERRORS OF METABOLISM IN BELGIUM

In Belgium, both the ‘Vlaamse Gemeenschap’ (VG, the Dutch speaking community) and the French speaking community (later called the ‘Fédération Wallonie-Bruxelles’ (FWB))[a] use the Guthrie test for their NBS programme. Since the early eighties when this responsibility was shifted from the federal level to the communities,, the list of diseases to be screened is decided by each community, according to their own legal criteria and upon guidance from their own steering committees. As a result, the two communities do not screen for the same set of diseases: currently 11 diseases are screened for in VG and 13 in FWB; 9 of these diseases are screened for in both communities (Table 1).

Table 1 – Diseases included in neonatal blood screening programmes at community level, as of February 2016

Disease	Abbreviation	In Vlaamse Gemeenschap	In Fédération Wallonie-Bruxelles	Included in this pilot study
Metabolic disorders				
Biotinidase deficiency	LMCD	Yes	No	**Yes**
Galactosemia	GAL	No	Yes	**Yes**
Glutaric acidemia type 1	GA I	Yes	Yes	
Homocystinuria	HCY	No	Yes	**Yes**
Isovaleric acidemia	IVA	Yes	Yes	
Leucinosis or Maple syrup urine disease	MSUD	Yes	Yes	
Medium chain acyl-CoA dehydrogenase deficiency	MCAD	Yes	Yes	
Methylmalonic acidemia	MMA	Yes	Yes	
Multiple acyl-CoA dehydrogenase deficiency	MADD	Yes	Yes	
Phenylketonuria	PKU	Yes	Yes	
Propionic acidemia	PA	Yes	Yes	
Tyrosinemia Type I	TYR I	No	Yes	**Yes**

[a] In FWB, the NBS programme is managed by the ‘Office de la Naissance et de L’Enfance’ (ONE) since 2015

Very long Chain CoA deshydrogenase deficiency	VLCAD	No	Yes	**Yes**
Endocrine disorders				
Congenital hypothyroidia	CHT	Yes	Yes	
Congenital adrenal hyperplasia	CAH	Yes	No	**Yes**

2.1 History of neonatal screening for metabolic diseases in Belgium

The first neonatal screening programme has been established in Belgium in 1968, with the systematic screening of all newborns for hyperphenylalaninemia/ phenylketonuria.[2] In 1974, five diseases have been added to the screening programme, tyrosinemia, leucinosis, histidinemia, homocystinuria and galactosaemia, followed in 1980 by hypothyroidia.[3] In the early eighties, the responsibility of this screening has been transferred to the Communities.[4]

2.2 Neonatal screening in the Fédération Wallonie-Bruxelles (FWB)

In the FWB, the original national legal text has been adapted with little changes in 2001. In 2009, a full FWB legal text (Arrêté) has been established, including a specific screening protocol. Three screening centres are recognized in FWB and ensure the screening of selected diseases of all newborns:[5]

- UCL : Centre de dépistage néonatal des Cliniques Universitaires St Luc, Cliniques universitaires St Luc
- ULB : Centre de dépistage néonatal de l'ULB, Université Libre de Bruxelles (HUDERF)
- ULg : Centre de dépistage néonatal de Liège, Centre Hospitalier Universitaire de Liège

In 2014, a new Arrêté has been published to adapt the list of diseases and practical organisation, and a detailed guideline on each disease and the testing procedures has been published.[6] Since late 2015, the neonatal screening is managed by the Office de la Naissance et de l'Enfant (ONE) following the State reform.

2.2.1 Organization of screening in the Fédération Wallonie-Bruxelles

The organisation of the screening programme in FWB is described in a protocol.[7] Around 60 000 newborns are screened every year in FWB, in one of the three screening centres. These centres carry out testing, inform maternities and independent midwives of the results, inform the reference physician of abnormal or positive tests and collect data. The screening cost is supported by the FWB administration. An amount of 12.35€ (2009 price, yearly indexed) by newborn is paid by the FWB to the screening centres.[5, 7] The families should not support any screening cost for these diseases.

Testing methods are similar across centres and described in the guideline.[6] If the first test is positive, the same test is usually repeated. For some diseases, a second-tier test is carried out, e.g. for tyrosinemia I (see Supplement). The referral physician is responsible for requesting confirmatory tests, which fall outside the screening programme (even if conducted in the same laboratories).

A Steering Committee has been established to provide technical support, and has the responsibility of proposing new diseases for screening, guide the programme, monitor quality indicators, and contribute to the sensitization of health professionals and institutions to neonatal screening.[5, 7]

2.2.2 Diseases targeted by newborn screening

As of January 2016, 13 diseases are included in FWB, among which are one endocrine and 12 metabolic diseases (Table 2).[7] In addition, the three reference centres also test for a number of additional diseases (through MS/MS), without funding from the FWB. For instance, the ULB and UCL reference centres also tests for citrulinemia, hyperornithinemia and long chain acyl-CoA dehydrogenase deficiency, and the UCL centre also test for argininemia.[8, 9] This list of diseases for screening is regularly updated, upon proposal of the Steering Committee. In 2014, seven diseases have been added to the initial list of six diseases (Table 2).

Table 2 – Inborn metabolic diseases screened in the Federation Wallonie – Bruxelles, over time

Disease	Abbrev.	2001	2009	2014
Hypothyroidia	CHT	Y	Y	Y
Leucinosis or Maple syrup urine disease	MSUD	Y	Y	Y
Homocystinuria (hypermethioninemia)	HCY	Y	Y	Y
Phenylketonuria	PKU	Y	Y	Y
Tyrosinemia type 1	TYR I	Y	Y	Y
Galactosaemia	GAL	Y	Y	Y
Very long-chain acyl-CoA dehydrogenase deficiency	VLCAD			Y
Medium chain acyl-CoA dehydrogenase deficiency	MCAD			Y
Multiple acyl-CoA dehydrogenase deficiency	MADD			Y
Methylmalonic acidemia	MMA			Y
Propionic acidemia	PA			Y
Glutaric aciduria type 1	GA I			Y
Isovaleric acidemia	IVA			Y

The following criteria are used by the Steering Committee to evaluate new diseases for inclusion:

1. The last Arrêté (2014) mentions two criteria: the evolution of scientific knowledge and a benefit-risk (avantages/inconvénients) analysis.[7]
2. The 2013 guideline describes a more detailed list of criteria (in Supplement), summarized below:[6]
   - o The disease must represent an important health problem, i.e. severe disease, with difficult early diagnosis and with irreversible

sequelae if left untreated. The epidemiology and disease course must be sufficiently known.

- The screening test must be simple, reliable, reproducible, scientifically validated and acceptable by the population.
- There is a consensus about confirmatory tests and further investigations of positive results.
- An effective intervention must exist for early detected patients, and evidence must show that early intervention is more effective than later intervention. A policy of case management and access to treatment is ensured for all patients.
- The effectiveness of the screening programme is evidenced by high quality randomized trials; or at least be supported by an international scientific consensus.
- The benefits of screening must exceed (by far) the inconveniences: test and their imprecisions, the diagnostic procedures and intervention.
- The financial resources are available and the programme costs must be comparable to other prevention interventions that are funded by public authorities for a similar result.

2.3 Screening of newborns born in Brussels

In principle, it is the language of the health care institution where delivery is taking place that determines whether a newborn will be under the FWB of the Flemish NBS programme.[10] In practice, this means that all newborns born in the UZ Brussel benefit from the VG programme, while those born in other Brussels hospitals are in the FWB programme, regardless of language of parents. In practice, samples of those born in the nine other Brussels hospital maternities are tested by the ULB or UCL reference centres (M Duguerry and T. Pereira, personal communication).

2.4 Neonatal screening in the Vlaamse Gemeenschap (VG)

From the early eighties up till 2002, the official screening programme in the VG consisted of testing for phenylketonuria and congenital hypothyroidia.[11-13] In January 2003 screening for congenital adrenal hyperplasia was added, followed in January 2007 by the screening for biotinidase deficiency and the introduction of tandem mass spectrometry (personal communication: F. Eyskens).

In 2003 the VG issued a decree defining the target areas of its preventive health policy as well as the general legal requirements for any programme active in these areas. A decision of the Flemish government followed then in 2008 to detail more specifically the criteria with which population-wide health screening programmes have to comply.[14, 15]

As a result neonatal screening, previously performed by up to five screening centres (one in each province), was reorganised in 2012 with only two screening centres remaining:[10]

- PCMA: Provinciaal Centrum voor opsporing van Metabole Aandoeningen
- VCBMA : Vlaams Centrum voor opsporing van aangeboren metabole aandoeningen, Universitair Ziekenhuis Brussel, Vrije Universiteit Brussel

The Agentschap Zorg en Gezondheid (AZG) is the designated authority to manage the neonatal screening programme and fixes the minimal requirements of the testing procedures and the organisational modalities of the programme.

2.4.1 Organization of screening in the VG

The organisation of the screening programme in VG is described in a protocol.[10,16] Around 70 000 newborns are screened every year in VG, through one of the two screening centres. These centres carry out testing, inform maternities and independent midwives of the results, inform the reference physician of abnormal or positive tests, collect data and promote the programme. The screening cost is supported by the AZG administration and amounts to € 21.19 per neonate screened (about € 1 500 000 yearly, indexed; personal communication: K. Colaert, AZG). The families do not have to pay for this screening.

Testing methods are similar across centres and described in the guidelines which are revised at regular intervals.[10, 16] If the first test is positive, the same test is usually repeated and/or a repeat sample may be requested depending on the condition being screened. The referral physician is responsible for requesting confirmatory tests, which fall outside the screening programme (even if conducted in the same laboratories).

Guidance is provided by a Steering Committee for population-wide health screening programmes which has the support of a separately nominated Flemish working group for population-wide screening on congenital disorders in neonates by means of a blood sample. The legal assignment of this last group consists specifically of: monitoring scientific developments and societal evolutions in the field of NBS; counselling on quality assurance and evaluation of the NBS programme by proposing criteria, indicators and improvements; recommending methods to raise awareness of the public, institutions and health professionals; work out evidence based proposals to include additional diseases in the NBS programme; and guiding the health administration on the organisational requisites of the NBS programme.[16]

2.4.2 Diseases targeted by newborn screening

As of January 2016, 11 diseases are included in VG (Table 2).[16] In addition, the two reference centres also test for a number of additional diseases (through MS/MS), without funding from the VG (e.g. very long-chain acyl-CoA dehydrogenase deficiency). This list of diseases for screening is regularly updated, upon proposal of the Flemish working group for population-wide screening on congenital disorders in neonates by means of a blood sample (Table 3).

Table 3 – Inborn metabolic diseases screened in the Vlaamse Gemeenschap, over time

Disease	Abbrev.	2010	2015
Hypothyroidia	CHT	Y	Y
Congenital adrenal hyperplasia	CAH	Y	Y
Leucinosis or Maple syrup urine disease	MSUD	Y	Y
Biotinidase deficiency	LMCD	Y	Y
Phenylketonuria	PKU	Y	Y
Medium chain acyl-CoA dehydrogenase deficiency	MCAD	Y	Y
Multiple acyl-CoA dehydrogenase deficiency	MADD	Y	Y
Methylmalonic academia	MMA	Y	Y
Propionic academia	PA	Y	Y
Glutaric aciduria type 1	GA I	Y	Y
Isovaleric academia	IVA	Y	Y

3 METHODOLOGY

3.1 Introduction

After a review of the literature and websites from selected countries we decided to base our methodology to evaluate diseases for newborn blood screening on an adaptation of the INESSS methodology (Institut National d'Excellence en Santé et en Services Sociaux, Québec, Canada) published in 2013.[18] The advantage of the INESSS method is that their report describes a clear methodology and its practical implementation for a number of inborn errors of metabolism. Other countries, for example the recent Dutch recommendations, rather used consensus meetings and public hearings as general features for decision making rather than a formalized method.[19]

3.2 Diseases included in this pilot testing

During the scoping of this study we agreed with the steering group to pilot test a Multi-Criteria Decision Analysis (MCDA, see below) approach on the *six diseases* that are screened for in only one part of the country. This selection of diseases by KCE and the steering group was based on pragmatic reasons with the aim to limit the number of diseases for this testing of a method but at the same time to make it relevant in the Belgian context.

Screened in VG only:

- Congenital Adrenal Hyperplasia (CAH)
- Biotinidase deficiency (LMCD)

Screened in FWB only:

- Galactosaemias (GAL)
- Tyrosinemia Type I (TYR I)
- Homocystinuria (HCY)
- Very long Chain CoA deshydrogenase defiency (VLCAD deficiency)

The potential inclusion of Cystic Fibrosis (CF) in this pilot test was discussed in the steering group. As no community in Belgium currently performs screening for this disease and the KCE published a report on CF screening in 2010 with valuable material readily available, it was decided not to include it. However, it could be an obvious candidate for the next round of evaluations.[20]

3.3 INESSS methodology

In short this methodology takes a two-step approach for a Multi-Criteria Decision Analysis (MCDA).[18] INESSS based this methodology on the EVIDEM framework, a freely available tool for decision analysis.[21]

Based on the proposed criteria in the EVIDEM framework the steering group at INESSS decided to use seven specific decision criteria for evaluating the advisability of NBS screening for each of the selected diseases. Selected diseases were based on the screening programme from the neighbouring Ontario province, later expanded with a few additional diseases. These criteria were provided with a relative weight for their importance by a twelve person steering group.

Those criteria are largely based on the original Wilson and Jungner criteria from 1968,[22] but adapted for rare diseases.

For each disease the available information to score the disease under evaluation was assessed. More information on the criteria and diseases can be found in the supplement (original French language version).

Disease information was retrieved through review of the literature and through registries (Quebec, Ontario, whole of Canada, Orphanet).[18] This information was then structured according to the selected criteria and presented in a comprehensive way in disease summaries for scoring them.

3.4 Adaptation of the INESSS methodology for this study

3.4.1 Working definitions

The following basic terminology was used throughout this report:

- **<u>Weights</u>: the general importance of the seven decision criteria selected (low to high: 1 - 4)**
- **<u>Scores</u>: the score given to a specific disease when considering these criteria (low to high: 0 - 4). Zero stands for 'unable to answer, missing data'. Those zero values are reported but excluded from the calculations.**

3.4.2 Steps in the study

In an MCDA a number of criteria are selected a priori and attributed a specific weigh. The diseases under assessment are subsequently scored on each of these criteria and ranked based on a global composite score including weight and score for each of the criteria.

The steps in this study are:

1. Select and define in more detail criteria for the evaluation of diseases;
2. Define weights by criterion that is later applied in the calculation of a global composite score by disease. These weights are expressed using a four point LIKERT scale (Table 4)
3. Prepare scientific information for each disease on all criteria in a short disease summary
4. Score each criterion, by disease. Also for these scores a four point LIKERT scale is used (Table 4)
5. Calculate a global composite score per disease: in general the weight for a criterion is multiplied with the score for a disease for that criterion. This weighted score is than summed per disease, leading to global composite score per disease or an individual composite score for each evaluator. In the last case de scores of the individual evaluators are aggregated to obtain a global composite score by disease, allowing there ranking.

Step 1 was conducted jointly by the KCE team together with the steering group consisting of decision makers from both communities, NBS laboratory experts, clinicians, patient representatives and an ethical expert (eleven evaluators in total). Steps 2 and 4 were done exclusively by the evaluators from the steering group while steps 3 and 5 were conducted by the KCE team and later discussed with the steering group.

The steering group and the KCE team jointly decided to select for this pilot study the six diseases that are screened in one community but not in the other (Table 1). Since the scope of this study is pilot testing a method, the results are not recommendations to include or exclude specific diseases from the NBS programs. The application of the methodology and the results allowed to learn a number of lessons that were discussed with the steering group evaluators, and that are described below together with the results.

3.4.3 Selection and definition of criteria

Similar criteria as used in the INESSS method were used and adapted, further defined and refined during the steering group meetings.

- Frequency (birth prevalence[b]) in Belgium and other Western countries.
- Severity of the disease in untreated cases
- Timely availability of the test results
- Efficacy of early treatment vs. late treatment
- Probability and impact of false positive results (negative impact)
- Probability and impact of false negative results (negative impact)
- Impact on the health care system

During the scoping exercise, it was first suggested to add other criteria, in particular legal requirements in one of the communities. A risk in having a long list of additional criteria is to have overlapping criteria and thus to run the risk of measuring the same aspect several times. If this happens it might lead to the probability that too much importance is given to that aspect, leading to imbalanced results at the end. Therefore, we further defined each criterion and detailed the specific aspects explicitly under each one of those seven general criteria (see 3.4.4).

After elaborate discussions, it was decided that ethical consideration should be included in the evaluation of all of the seven criteria selected, without being a separate criterion.

3.4.4 Content of each of the criteria

To define more precisely what is included in each of these criteria, we interpreted these criteria for our Belgian context as follows.

1. Frequency (birth prevalence) in this country/this community, including

[b] Incidence and prevalence are not always clear in the definitions of newborn screening. As a working definition we will use the term frequency to indicate *'Birth Prevalence at live birth detected through newborn screening'*. Prevalence in the population might be different due to differences in survival or how severity of the disease should be judged. Further information about this ongoing discussion on terms can be found in Mason et al.[23]

   - Where available birth prevalence in Belgium
   - Birth prevalence in Western Europe, North America and world-wide
2. Severity of the disease in untreated cases, includes
   - Description and severity of the disease when untreated, mortality, QALY loss due to the disease
   - Epidemiology and natural history of the disease
3. Timely availability of the test results, includes
   - Is the test result available at a timely moment to avoid preventable complications and sequelae before a diagnosis based on suggestive clinical signs would have been made (related to disease spontaneous evolution)
4. Efficacy of early treatment vs. late treatment, according to type of treatment (specific, non-specific), includes
   - What is the efficacy of early treatment compared to later treatment
   - What is the evidence that participants will probably benefit from this screening and not be harmed by it, for example through early diagnosis of an untreatable disease that will only appear at a later age (except for those related to false positive and false negative)
   - Consensus about the diagnostic pathway? Circuit after a test-positive result
   - Consensus about management of the disease if confirmed
   - Availability of diagnostic and disease management facilities
5. Probability and impact of false positive results (negative impact), includes
   - Are FP reported and how frequent are they, if not reported how probable would they be
   - Impact of false positive results
6. Probability and impact of false negative results (negative impact), includes
   - Are FN reported and how frequent are they, if not reported how probable would they be
   - Impact of false negative results
7. Impact on the health care system, includes
   - Impact on diagnostic capability and treatment capability, including management of detected (true) case and of side diagnoses
   - Cost of (adding) a specific disease to the existing programme
   - Diagnostic cost for confirmatory tests when screening is positive (both TP and FP)
   - Cost of case management of confirmed disease (care payer and societal)
   - Cost-effectiveness of screening for this disease when available
   - Communication towards participants about potential benefits and harms
   - Organizational aspects of screening and its feasibility
   - Acceptability of diagnosis strategy by the population and patient satisfaction

3.4.5 *Weighing of the criteria*

The relative weight of each criterion was calculated based on expert opinion from the steering group (see below) the result of which was afterwards submitted to a consensus of the participants. Reweighting after further detailing of the content of each of the criteria was performed after a second expert meeting.

Weighing of these criteria was done by asking the steering group to score criteria using a LIKERT scale. To avoid neutral scores, and based on the advices of experts in the field of semi-quantitative and qualitative research, and the EVIDEM framework,[21] we opted for a scale from 1 to 4 to score criteria.

Table 4 – Description of the LIKERT scales used for scoring criteria weight and disease

Score	Criteria (for weight)	Score of disease for this criterion
1	not relevant	very low
2	slightly relevant	rather low
3	relevant	rather high
4	extremely relevant	very high
0		unable to answer, missing data

The participating steering group members received Excel sheets to weigh those criteria and if they wanted to comment on their opinion.

From these individual LIKERT scale weights both the mean and the median absolute scores are calculated.

Based on internal discussions with other KCE researchers and previous MCDA approaches across the world,[18, 24, 25] and for consistency among projects we decided for the main analysis of the global composite scores to use a common proportional (median) weight for the criterion on a percentage scale for all seven criteria (see Equation 1).

Equation 1 – Calculation of Median Proportional Weight (weights for seven criteria)

$$P\ Median\,Weight_{Criterion\ x} = \frac{Median\,Weight_{Criterion\ x}}{\sum_{i=1}^{7} Median\,Weight_{Criterion\ i}} \quad * \quad 100\%$$

P Median Weight: Proportional Median Weight. Median Weight: Median weight on a 1 - 4.weight LIKERT scale.

In the analysis we also present a stratified analysis by type of evaluator (labs vs. other) to compare their weighing of criteria.

3.4.6 Preparation of disease-specific scientific information

For each of the six selected diseases, information on each criterion was synthesised into a disease summary (see supplement). Information on those diseases was collected from the participants of the steering group, the Belgian data from the communities and completed with a pragmatic literature review, and information from other screening and research agencies. However, information was not always readily available due to the low frequency of occurrence of these diseases.

To estimate the impact of detected cases on the health system we recalculated the estimated birth prevalence in Belgium using the average of 125 000 live births in Belgium in recent years. This number of 125 000 live births was derived from the average life births recorded in statistics 2011-2014 from Statistics Belgium.[26]

3.4.6.1 Source of information and general data sources

The most important sources of disease information are listed below, and additional sources, including published studies, are referenced in the disease summaries in the supplement.

Flanders: Vlaams bevolkingsonderzoek naar aangeboren aandoeningen bij pasgeborenen via een bloedstaal. Draaiboek 2012 and 2015.[10, 16]

Wallonia: Guide pour le programme : Le dépistage néonatal des anomalies métaboliques en Fédération Wallonie-Bruxelles.[6]

The Netherlands Gezondheidsraad: Gezondheidsraad. Neonatale screening: nieuwe aanbevelingen.[19]

Orphanet[27]

INESSS report Quebec published in 2013.[18]

Ontario screening programme.[28]

Journal of Inherited Metabolic Diseases: J Inherit Metab Dis

The Online Metabolic and Molecular Bases of Inherited Disease.[29]

National Institutes of Health: Genetics Home reference.[30]

National Institutes of Health: Genetics Home reference.[30]

Additional sources: grey literature sources including other agencies, patient support groups, Google and Wikipedia were searched and any retrieved information subsequently validated through literature searches.

3.4.7 Scoring diseases for each criterion

As explained under 3.2, for this pilot assessment we choose to use the six inborn error diseases that are screened in one part and not in the other part of Belgium as test diseases.

3.4.7.1 Scoring method

We scored the diseases using a 4 point LIKERT scale (for the same reason as for choosing a 4 point scale for weighing the criteria, see above), for each criterion and each disease, based on information included in the disease summary (see above). To avoid misunderstanding in the scoring of diseases, we used the same LIKERT scale throughout the exercise, going from very low to very high. This means that in some cases we needed to convert the direction of the scale during the analysis. For example the impact of incidence/prevalence being very high might be an argument in favour of screening, while the impact of false positives being high might be an argument in disfavour of screening. After a preliminary scoring, an evaluation meeting was organized to agree on the face validity of the results. During this evaluation meeting participants could adapt their score when there appeared to be misunderstandings about the exact meaning of the weights and the scores.

The scores for each of the criteria:

- 1: Very low
- 2: Rather Low
- 3: Rather High
- 4: Very High

After completing the disease summaries we realised that not we did not have data to answer all criteria satisfactorily. Therefore, we added a last option to opt out for scoring a criterion for the specific disease:

- 0: unable to answer, missing data

In the final analysis we corrected for missing values by eliminating the specific criterion.

3.4.8 Guidelines for scoring

During the preparatory meetings several options to harmonise scoring were discussed. Ranges of quantitative values were pre-defined for two quantitative criteria, to facilitate a homogenous scoring of these criteria across participants.

Frequency of the disease (birth prevalence as detected by screening)

- <1 / 100 000: score 1
- 1 to 5 /100 000: score 2
- >5 to 10 / 100 000: score 3
- >10 / 100 000: score 4

Due to the uncertainty of the estimated birth prevalence, different scores were however allowed, due to personal opinion and experience of the participant who scores.

Timely availability of test results.

The same approach was proposed: the number is the proportion of results available on time to prevent serious complications.

- < 50% timely availability of results: score 1
- 50 to <80 % timely availability of results: score 2
- 80 to <100 % timely availability of results: score 3
- 100% timely availability of results: score 4

Again, due to the uncertainty of the estimated timely availability, different scores were allowed due to personal opinion and experience of the person who scores.

3.4.9 Calculation of global composite scores by disease

In all analyses we correct for missing values (scores zero) for a specific criterion by eliminating the scores for those specific criteria. For the last three criteria (probability and impact of false positives, of false negatives, and the impact on the health care system) we inverted the scores during the analysis to calculate global composite scores for each disease.

3.4.9.1 Main analysis

In this analysis, we considered weights and scores as ordinal values. For methodological reasons we preferred to use medians rather than means. We present for each disease the median of the sums of the products of the proportional median weight (by criterion) by the personal scores of each evaluator (see Equation 2). We also perform limited stratified analyses by type of participant to evaluate the impact of the participant profile.

Equation 2 – Calculation of the global composite score by disease (main analysis)

$$Median_{\text{ocross all respondents}}\left(\sum_{i=1}^{7} P\ Median\,Weight_i * IS_i\right)$$

P Median Weight: Proportional Median Weight. Median Weight: Median weight on a 1 - 4.weight LIKERT scale; IS: Individual disease score.

3.4.9.2 Additional analyses

There is no gold standard on how to construct these composite scores. With the aim to test whether other methods would give more divergent scores and for testing the stability of the analysis we tested five additional methods for analysing the same data. For these analyses we used the absolute weights for criteria and scores for diseases (on a LIKERT scale of 1 to 4, excluding the zeros from the scores but adjusting for the missing values in the summation).

In the description below, mean (median) common weight for each criterion stands for mean (median) weight for this criterion across all evaluators (abbreviation MW or MedW). Similarly, mean (median) score for each criterion by disease stands for mean (median) score across all evaluators for this criterion (abbreviation MS or MedS). In methods 3 to 6 we also used individual composite score calculations using individual weights (IW) and individual disease scores (IS).

Sum stands for the summation of the seven resulting composite scores (weight * score) by criterion for methods 1 and 2, and from the individual composites scores by evaluator for methods 3 to 6.

For each of the six diseases we used the following six methods. We then compared the ranking of disease between those six methods.

1. Sum (over all seven criteria) of mean common weight across evaluators multiplied by the mean score across evaluators: $\sum_{i=1}^{7} MW_i * \mathrm{MS}_i$
2. Sum (over all seven criteria) of median common weight across evaluators multiplied by the median score across evaluators: $\sum_{i=1}^{7} MedW_i * \mathrm{MedS}_i$
3. Mean of the weighted summed score per evaluator over all seven criteria of the individual weights multiplied by the individual scores given by each evaluator. The weighted summed score per evaluator is given by: $\left(\sum_{i=1}^{7} IW_i * IS_i\right)$.
4. Median of the weighted summed score per evaluator over all seven criteria of the individual weights multiplied by the individual scores given by each evaluator. The weighted summed score per evaluator is given by the same formula as in point 3.
5. Mean of the weighted summed score per evaluator over all seven criteria of the common median weight across evaluators multiplied by the individual scores: The weighted summed score per evaluator is given by: $\left(\sum_{i=1}^{7} MedW_i * IS_i\right)$.
6. Median of the weighted summed score per evaluator over all seven criteria of the common median weight across evaluators multiplied by the individual scores. The weighted summed score per evaluator is given by the same formula as in point 5. This method provides the same ranking as the main analysis in this report.

4 RESULTS OF CRITERIA WEIGHING AND DISEASE SCORING

4.1 Introduction

After a scoping meeting and several subsequent expert meetings we performed the final weighing of criteria and the scoring of the six selected diseases at the end of 2015. Our steering group represented eleven different organisations: the five Belgian centres for neonatal screening (the ‘*labs*’), two from the Belgian decision makers (the ‘*communities*’) and four from a ‘*patient perspective*’ (academic ethicist, patient group, a clinical expert and Orphanet Belgium).

Most centres sent in only one response for their centre, however two centres did provide exactly similar weights and scores from two members of the same organisation. Therefore, we excluded exactly duplicate responses from the analyses to avoid double counting. This way we obtained eleven single responses.

To analyse whether evaluators from different groups provided fundamentally different answers we also performed a limited stratified analysis by type of evaluator (five vs. six for the labs and stakeholders respectively, total n=11). Additionally, we performed a stratified analysis by language groups. Since one person explicitly works for both communities these answers were counted in both arms of the stratified analysis (seven for Dutch and five for French speaking community, total n=12). Moreover, this language distinction is not always clear since some academic and patient representatives effectively work for both communities.

4.2 Final steering group meeting

In January 2016 there was a final meeting to discuss the details of the weighing and scoring and to discuss potential misunderstandings. Misunderstanding about the direction of the scores by disease were raised and the analyses were corrected accordingly.

During this meeting, participants from the steering group were provided with their initial responses and were able to adapt their scores after further clarification and in case of misunderstandings. No final ranking of diseases was provided before the start of the meeting with the aim to avoid bias in adapting the personal scores. During the meeting, four of the evaluators slightly changed their scores.

This chapter gives an overview of the results of the weighing and scoring, as well as the ranking results using the main method of analysis for combining those. The alternative methods of ranking are also presented shortly. More details are given in the supplement.

4.3 Weighing of criteria

4.3.1 Overall weights

Each evaluator weighed the importance of each of the seven decision criteria. The means and medians of absolute scores are shown in Figure 2. The figures show that differences between medians and means of each criteria are often small, but the range of the overall criteria weight is wider when using the mean (2.4 to 4.0) than when using the median (3 to 4). The proportional weight as described in the methods section is shown in Figure 3.

The mean and median weight given to each of the criteria by the participants do not show substantial variations across criteria (absolute LIKERT scale weights in Figure 2 and proportional weights in Figure 3) but a trend for higher weight is observed for the criteria disease severity, efficacy of early treatment vs. late treatment and the probability and impact of false negative results (missed cases).

However, these means and medians hide the fact that the distribution of weights among evaluators is heterogeneous, as could be expected from a group composed with different profiles (Figure 4). This figure shows that the maximum score was given by all experts to disease severity. Unimodal distributions of the weights are seen for efficacy of early vs. late treatment, probability and impact of false positive results and the impact on the health care system. Higher dispersion across individual weights (bimodal or flat distribution) are observed for these three criteria: disease frequency, timely availability of test results and the probability and impact of false negative results. Those differences were discussed during the final expert meeting in January 2016 to solve potential misunderstandings about interpretation of questions.

Figure 2 – Weights (means and medians) for the seven criteria (on a LIKERT Scale from 1 to 4)

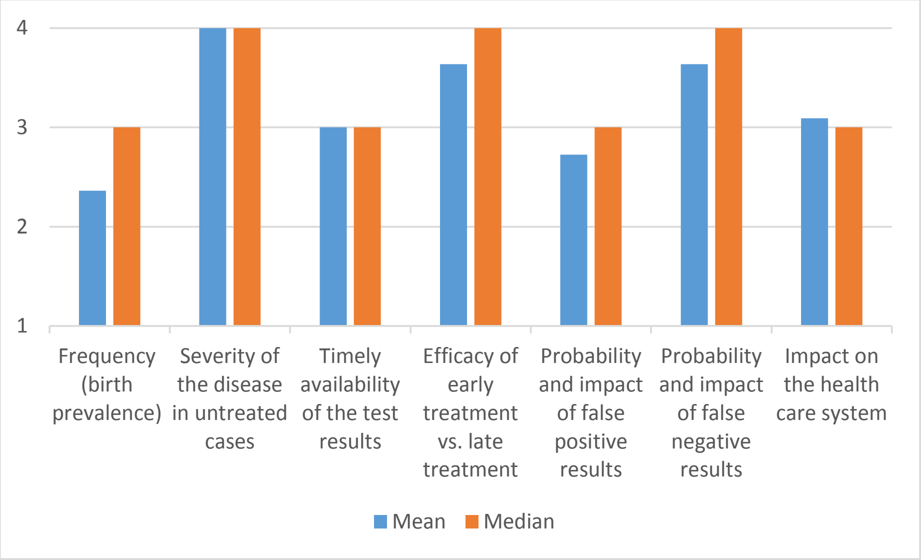


Figure 3 – Proportional weights (means and medians) on a global scale of 100% for the seven criteria

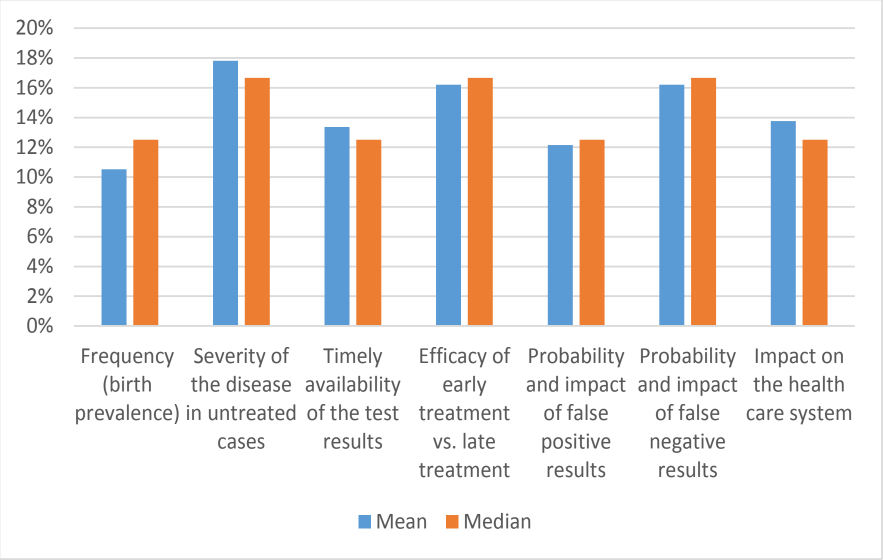
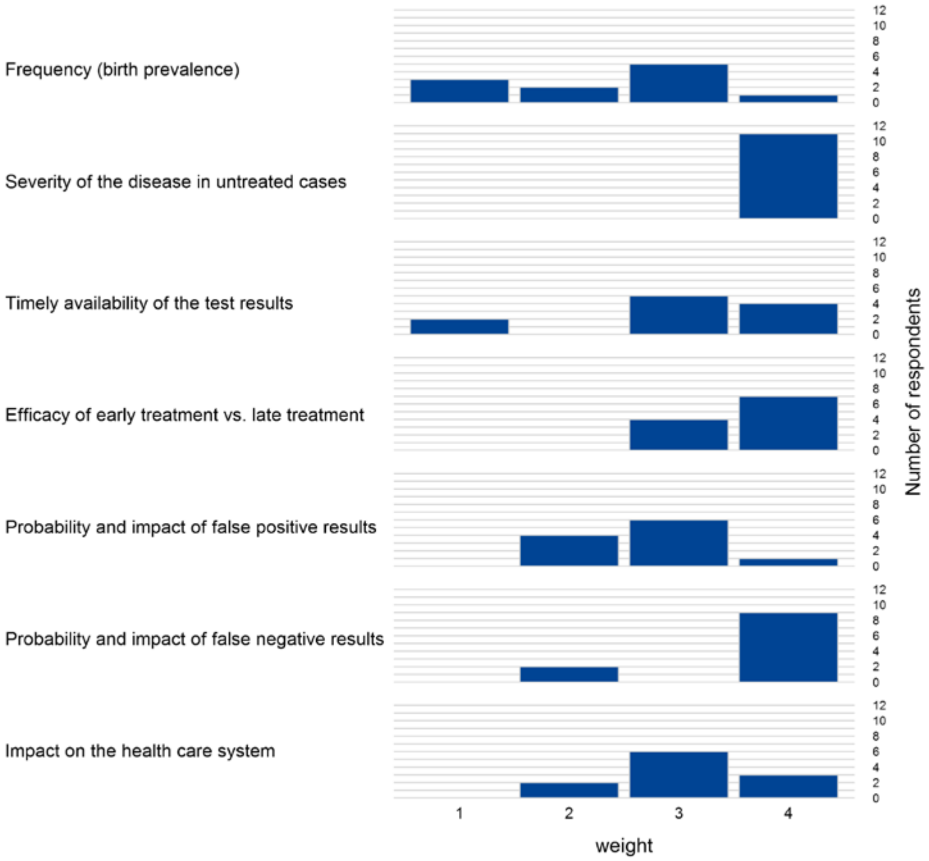


Figure 4 – Distribution of the weight of criteria given by participants

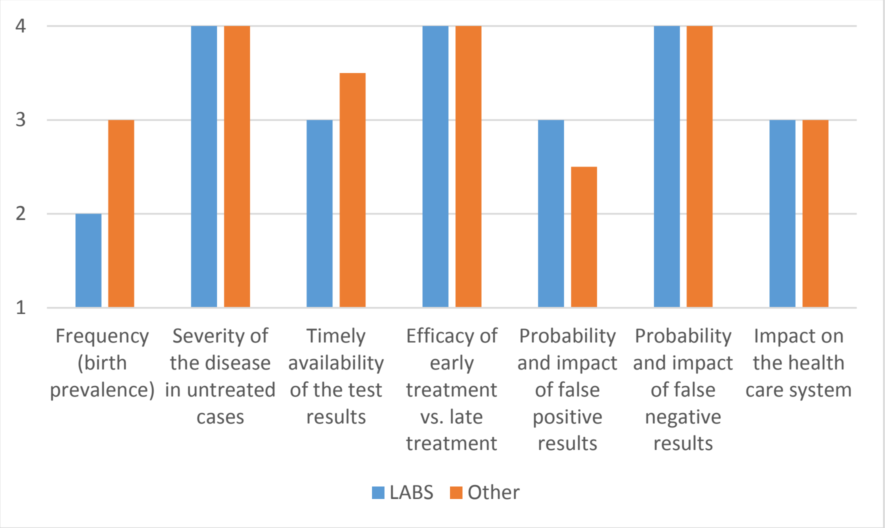


4.3.2 *Stratified weights: labs vs. other*

To detect whether the background of the evaluator providing weights for those criteria is important we performed a limited stratified analysis. Since we only had eleven evaluators this stratified analysis has limitations and was limited to only two categories: laboratory experts performing the NBS analyses (n=5) and non laboratory evaluators (n=6).

From Figure 5 is appears that the median weights for the criteria frequency and timely availability of test results are lower for the lab evaluators but that the criterion probability and impact of false positive results is scored somewhat higher by them. However, this stratified analysis is done on very small numbers so it is not very reliable.

Figure 5 – Stratified analysis of median weights attributed by lab evaluators vs. other evaluators



4.4 Scores by disease

After weighing the criteria, evaluators scored the six diseases (on a LIKERT scale 0 - 4) for each of the 7 criteria, leading to a total of 42 disease scores.

In the figures below we show the unweighted mean and median of those scores and the distributions of the scores by participant, for each criterion, by disease. As explained in the methodology, scores zero (unable to answer because of missing data) were excluded from the analysis, but the zero answers are described below each figure.

4.4.1 Biotinidase deficiency (LMCD)

Biotinidase deficiency received the highest scores for the criteria efficacy of early vs late treatment and timely availability of test results (Figure 6). It received a lower score (around 2) for frequency, probability and impact of both false positive and negative results. The distribution of individual scores by criterion (Figure 7) shows a high dispersion of individual answers, except for the criterion disease severity.

Figure 6 – Unweighted mean and median of scores for biotinidase deficiency for each of the criteria (means and medians)

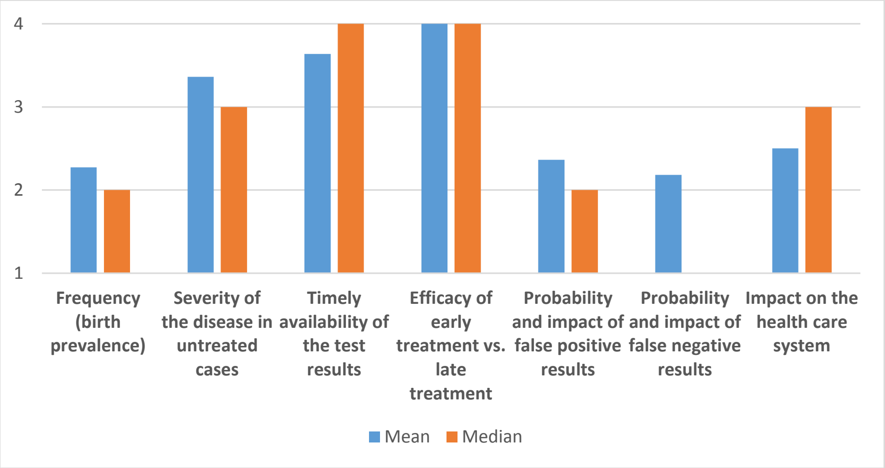
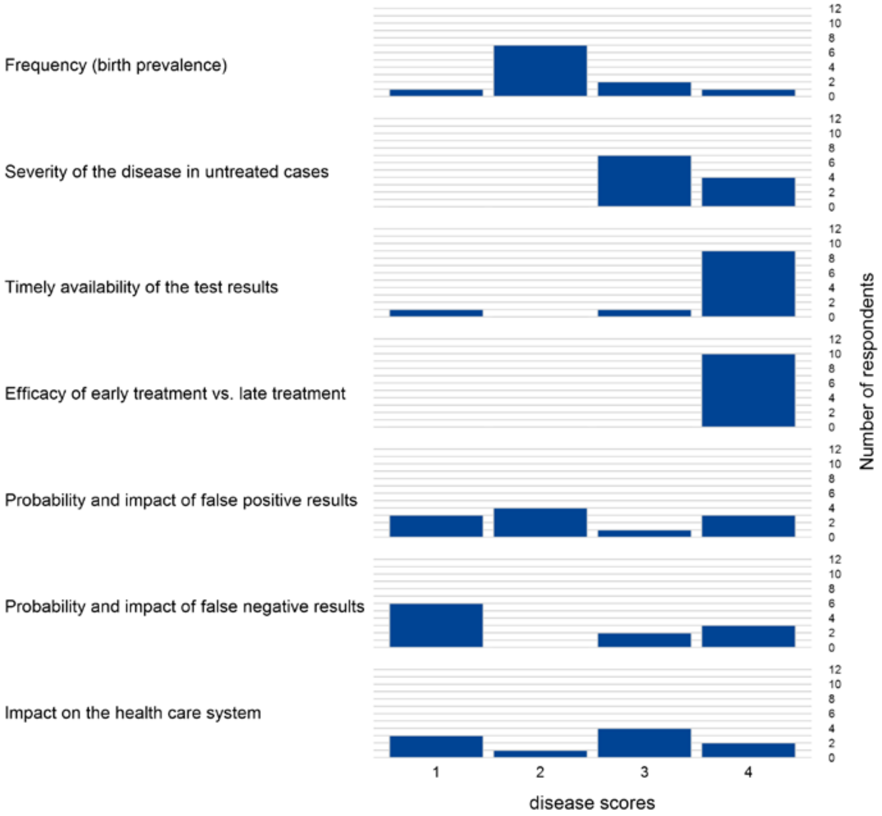


Figure 7 – Distribution of the disease scores for biotinidase deficiency



Not enough data to score was recorded for one evaluator for efficacy of early treatment and for another evaluator for impact on the health care system.

4.4.2 Congenital Adrenal Hyperplasia (CAH)

CAH received the highest score for the criteria efficacy of early vs late treatment and a low median score for the probability and impact of false negative results (Figure 8). The distribution of individual scores by criteria (Figure 9) shows a high dispersion of individual answers for all criteria, in particular for disease frequency.

Figure 8 – Unweighted mean and median of scores for Congenital Adrenal Hyperplasia for each of the criteria (means and medians)

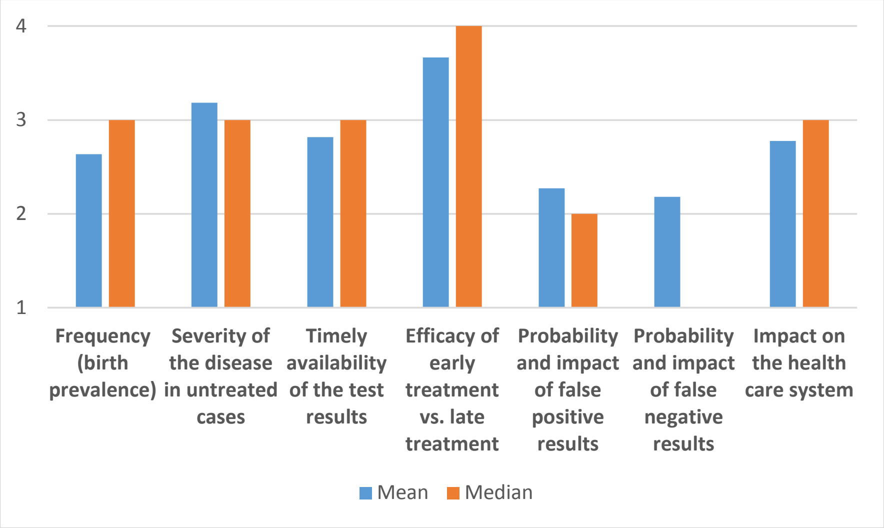
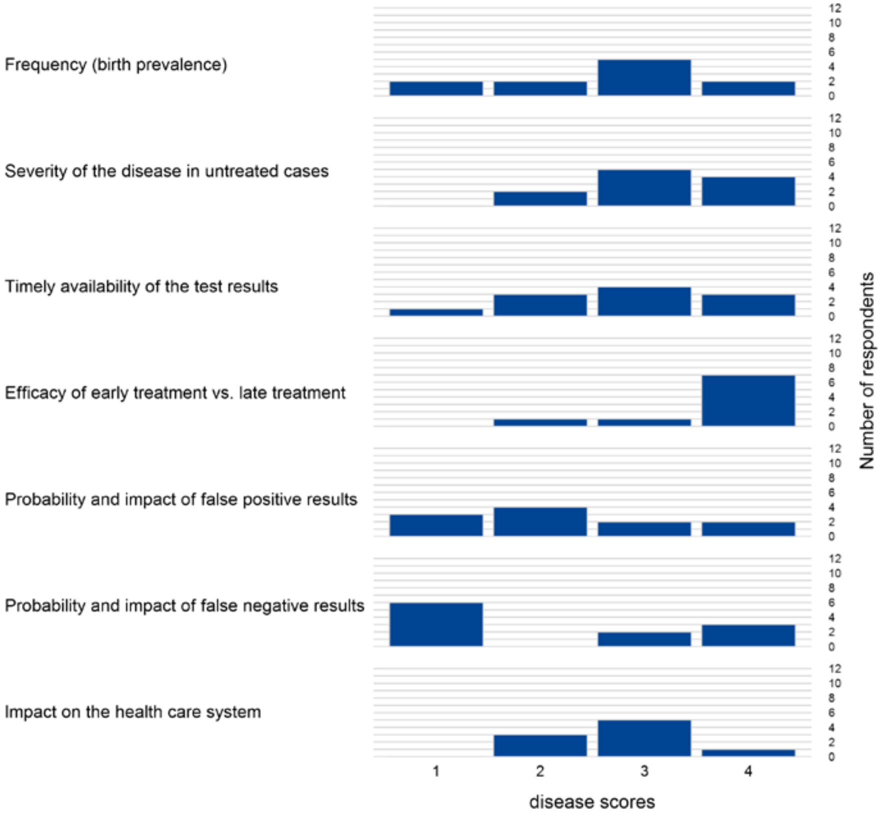


Figure 9 – Distribution of the disease scores for Congenital Adrenal Hyperplasia



Not enough data to score was recorded for two evaluators for efficacy of early treatment and for two other evaluators for impact on the health care system.

4.4.3 Galactosaemias (GAL)

GAL received the highest score for the criteria disease severity and lower median scores (2) for the disease frequency, probability and impact of false negative and positive results (Figure 10). The distribution of individual scores by criteria (Figure 11) shows a high dispersion of individual answers for all criteria, except for disease severity.

Figure 10 – Unweighted mean and median of scores for galactosaemias for each of the criteria (means and medians)

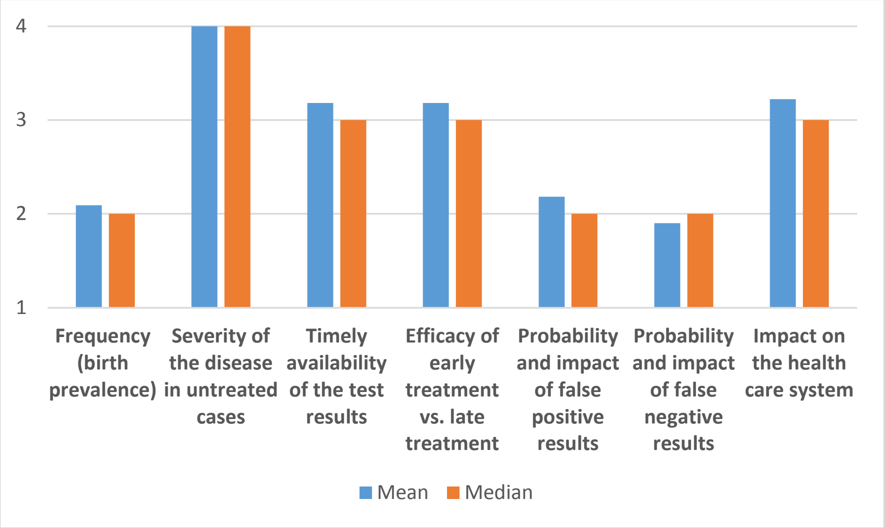
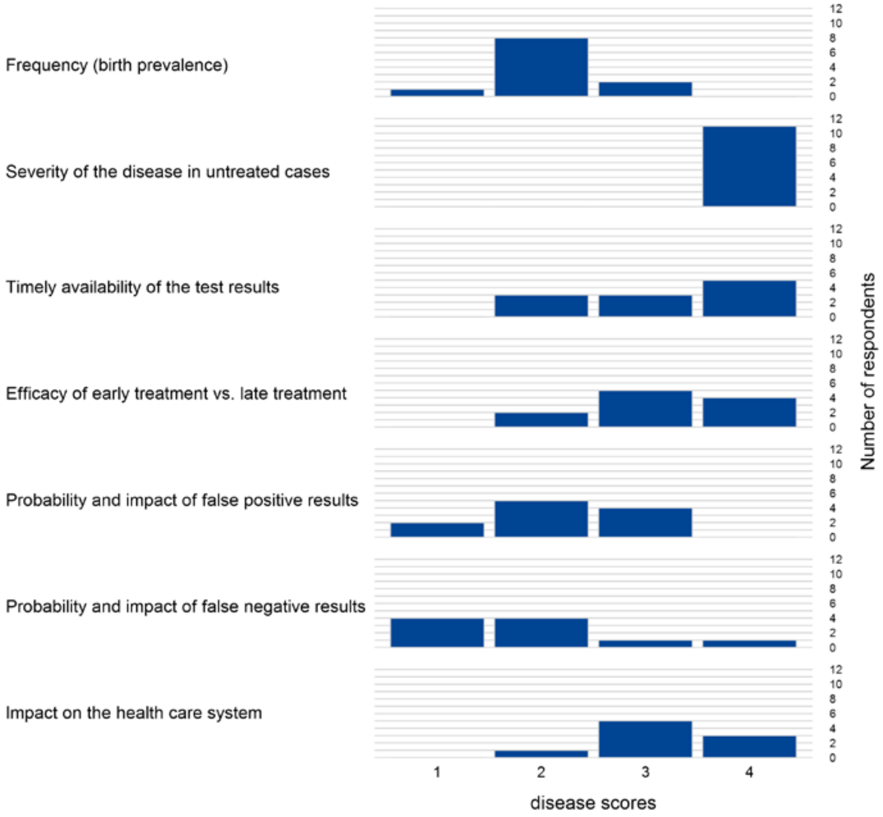


Figure 11 – Distribution of the disease scores for galactosaemias



Not enough data to score was recorded for one evaluator for false negatives and for impact on the health care system and for one additional evaluator for impact on the health care system only.

4.4.4 Homocystinuria (HCY)

HCY received the highest score for the criteria disease severity and timely availability of test results and very low median scores (1) for the disease frequency and probability and impact of false negative results (Figure 12). The distribution of individual scores by criteria (Figure 13) shows a high dispersion of individual answers for most criteria, except for disease frequency (but one outlier) and disease severity. Very divergent answers, with bimodal distribution at the lowest and highest scores (1 and 4) are observed for the criteria probability and impact of false positive results and impact on the health care system, which are not by nature reflected by the mean score.

Figure 12 – Unweighted mean and median of scores for homocystinuria for each of the criteria (means and medians)

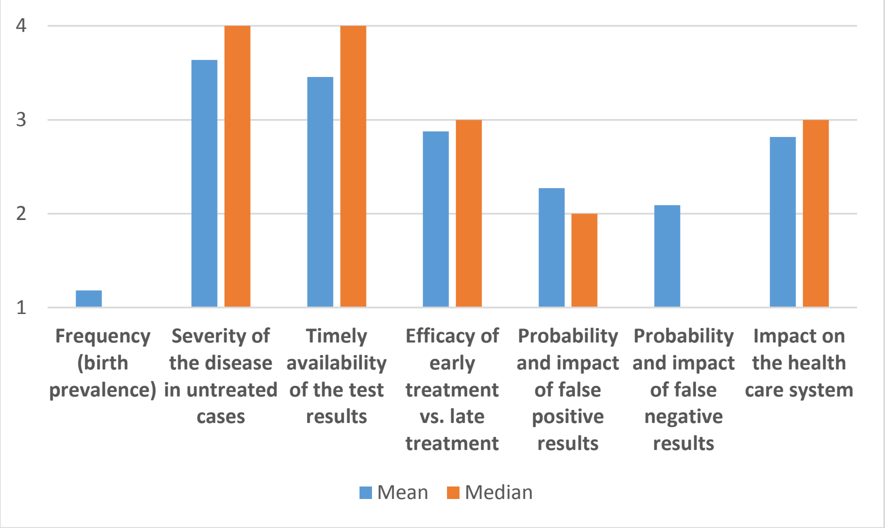
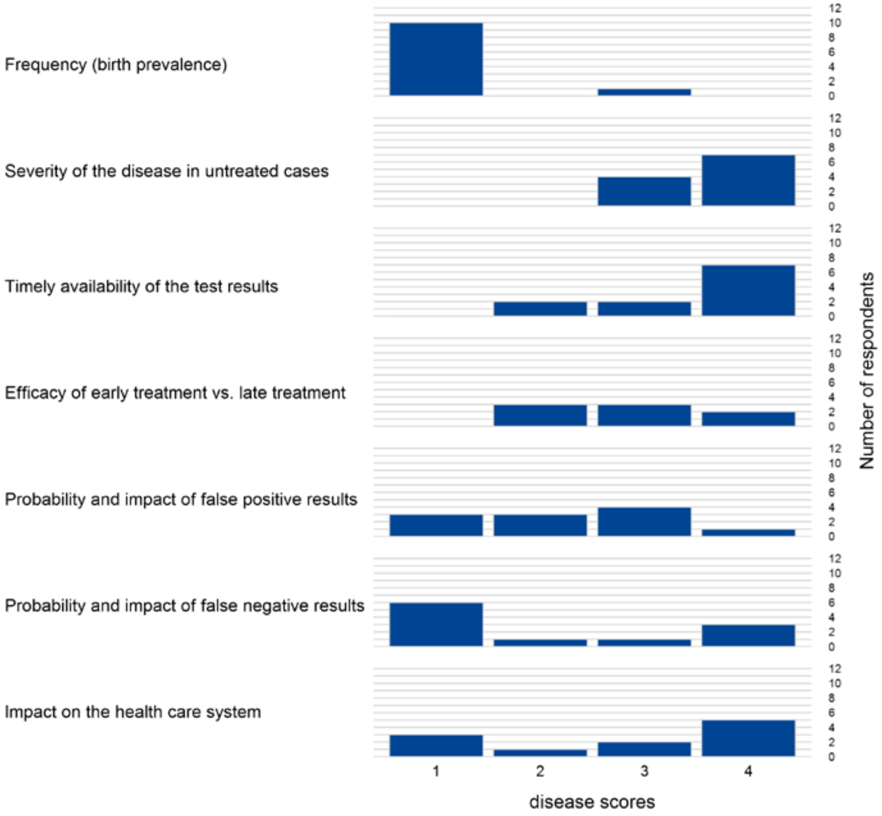


Figure 13 – Distribution of the disease scores for homocystinuria



Not enough data to score was recorded for three evaluators for efficacy of early treatment.

4.4.5 Tyrosinemia type I (TYR I)

TYR I received high scores (median at 4) for three criteria: disease severity, timely availability of test results and efficacy of early vs. late treatment. Low median scores (2) were given to disease frequency and probability and impact of false negative results (Figure 14). The distribution of individual scores by criteria (Figure 15) shows high dispersion of individual answers for two criteria only, related to false positive and negative results.

Figure 14 – Unweighted mean and median of scores for tyrosinemia type I for each of the criteria (means and medians)

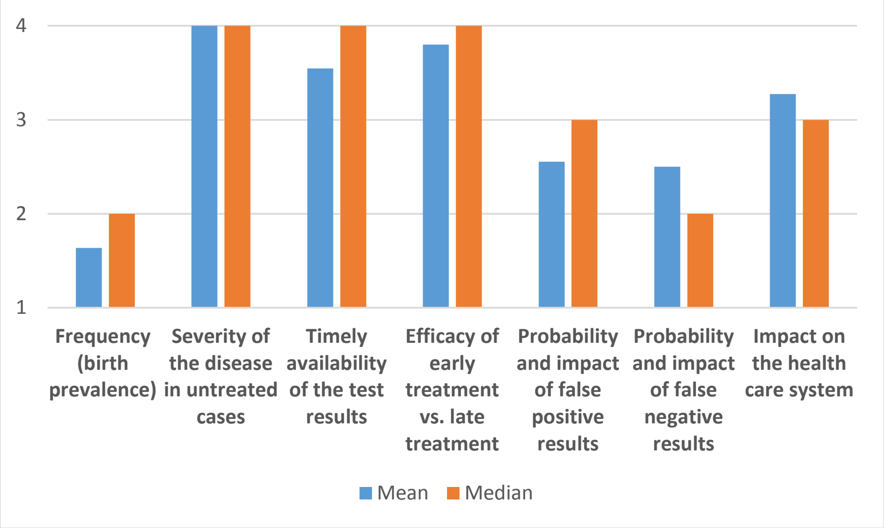
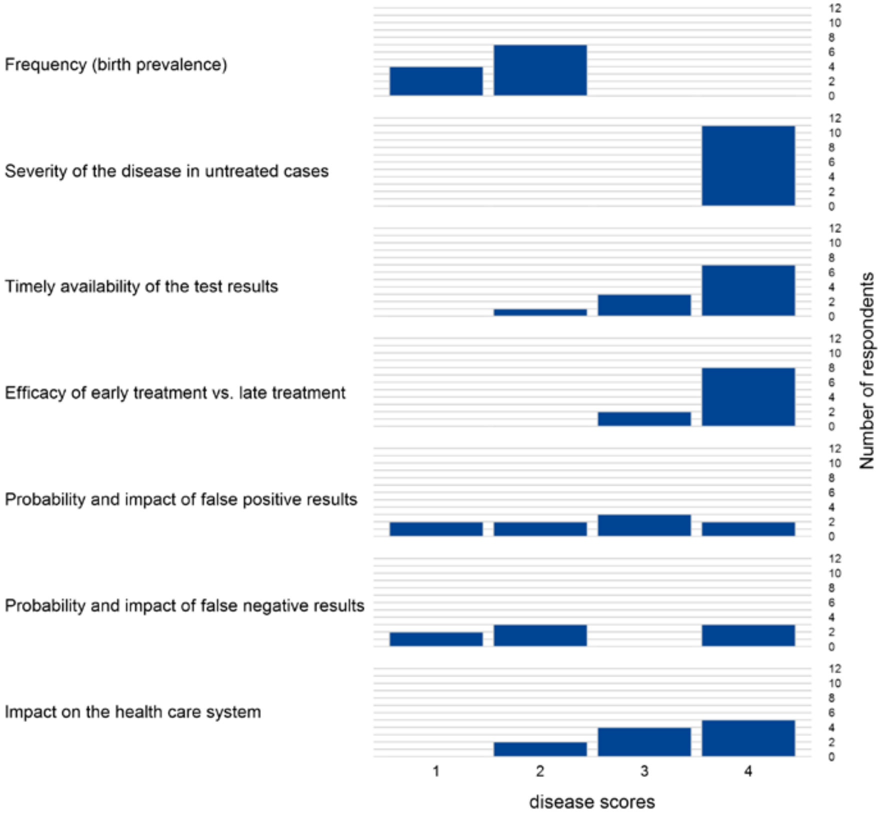


Figure 15 – Distribution of the disease scores for tyrosinemia type I



Not enough data to score was recorded for two evaluators for efficacy of early treatment. Those same evaluators also recorded “not enough data” for two other criteria, one for impact of false positives, the other for impact of false negatives.

4.4.6 *Very long chain Acyl CoA dehydrogenase deficiency*

Very long chain Acyl CoA dehydrogenase deficiency received a high score (median at 4) for disease severity and a very low score (median at 1) for disease frequency (Figure 16). The five other criteria achieved scores between 2 and 3 but the distribution of individual scores by criteria (Figure 17) shows a high level of dispersion for all of them. These distributions may question the validity of relying on a central tendency measure.

Figure 16 – Unweighted mean and median of scores for Very long chain Acyl CoA dehydrogenase deficiency for each of the criteria (means and medians)

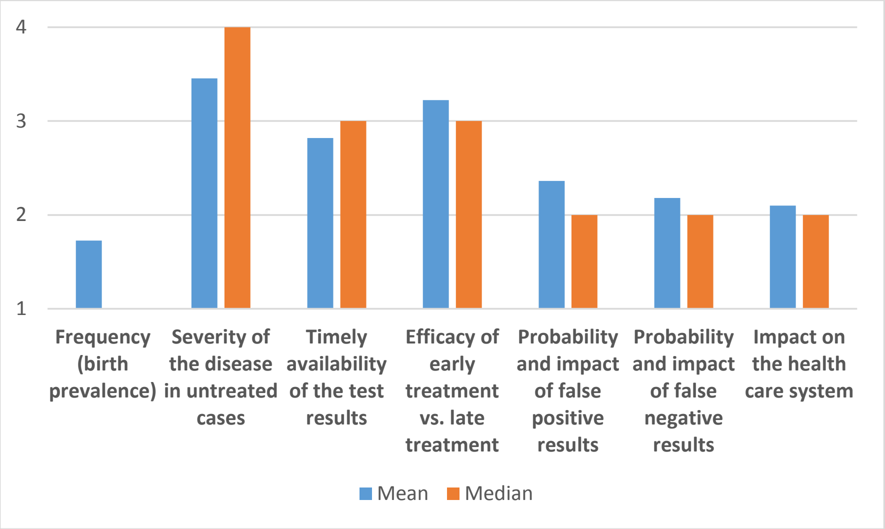
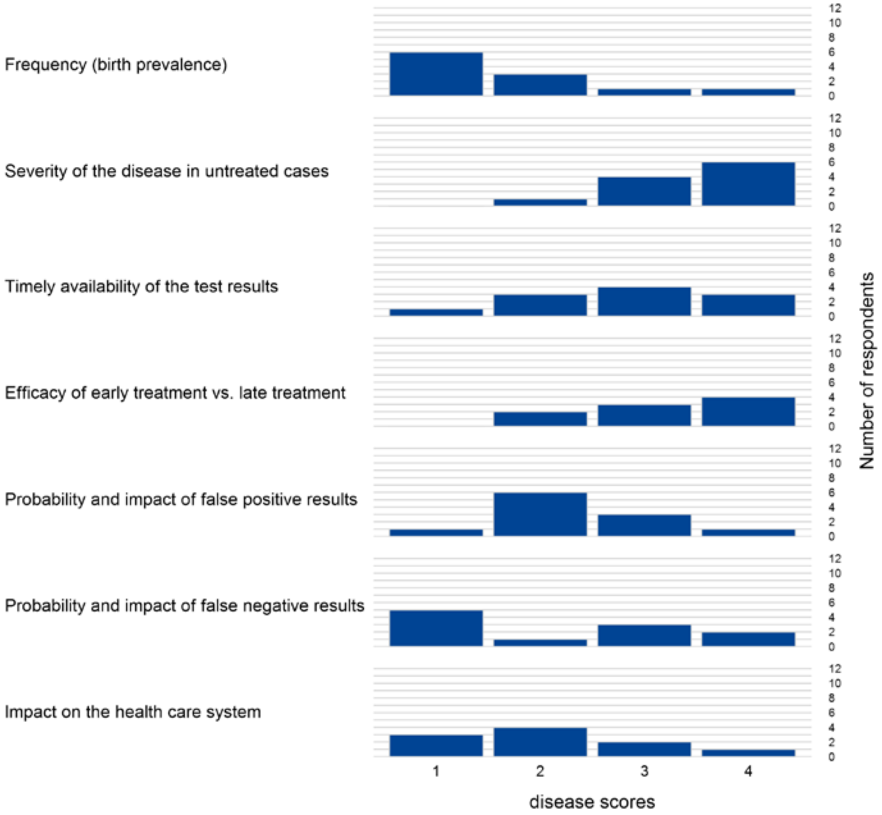


Figure 17 – Distribution of the disease scores for Very long chain Acyl CoA dehydrogenase deficiency



Not enough data to score was recorded for two evaluators for efficacy of early treatment. Additionally one those evaluators recorded not enough data for the impact on the health care system.

4.5 Composite scores

4.5.1 Global composite scores: main analysis

The previous data show that the differences in disease scores are not very substantial. As a consequence, the global composite scores also show little difference across diseases as can be seen in Figure 18. The median composite scores range from 2.9 to 3.1 but with a large range of individual composite score by evaluator.

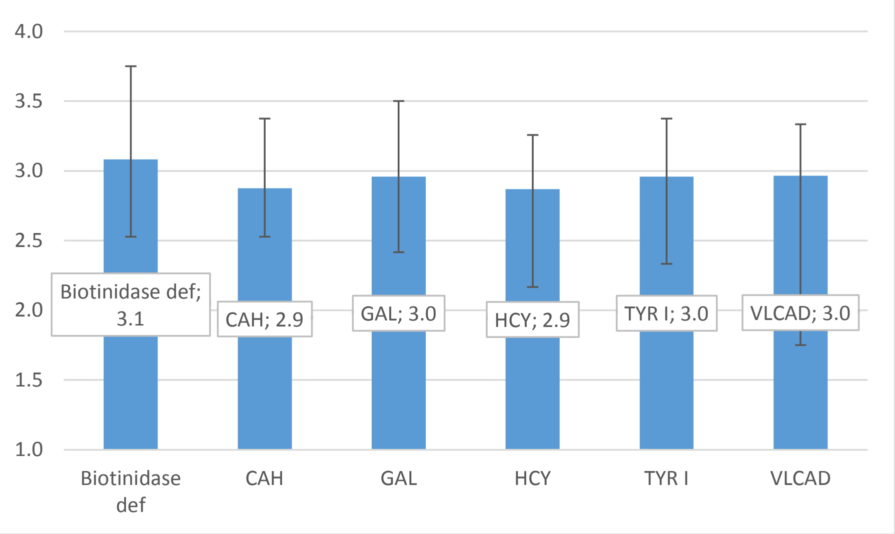
In this main analysis, we multiplied the global median proportional weight for each criterion for all evaluators with the individual scores per disease from each evaluator and summed it to an individual composite score (see methodology section). From these summed individual composite scores the median is taken as an aggregated global composite score for the whole group.

Additional results from five additional methods to obtain composite scores (see methodology section) are shown at the end of this chapter (section 4.5.4.).

These results by disease in the main analysis are very close to each other and the low potential to discriminate between diseases may seem disappointing at first glance.

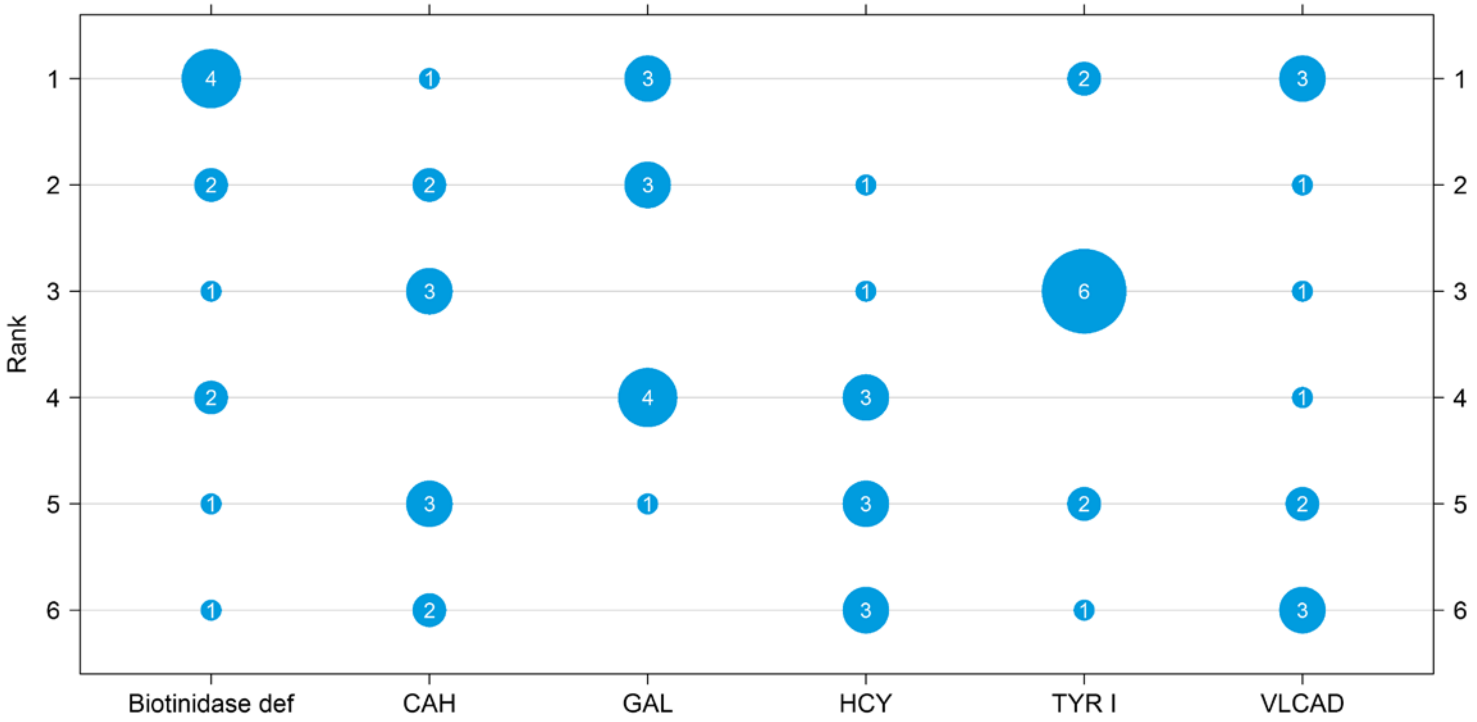
However, using the ranking by evaluator (in which rank 1 means ranked first and subsequently) gives a somewhat clearer view on the agreement and disagreement about ranking specific diseases between evaluators (see Figure 19). For example biotinidase deficiency is ranked first or second by over 50% of evaluators, while HCY is never ranked first and ranked second by only one evaluator. This method should therefore be able to provide some help in the ranking of diseases, not to represent a final ranking on itself but to allow a more formal discussion between decision makers during the decision making process.

Figure 18 – Base case analysis: composite scores by disease (+ range of individual composite scores by evaluator, n=11)



Median of the sum of the products of the median proportional weight by the personal scores (plus range of weighed scores) of each evaluator.

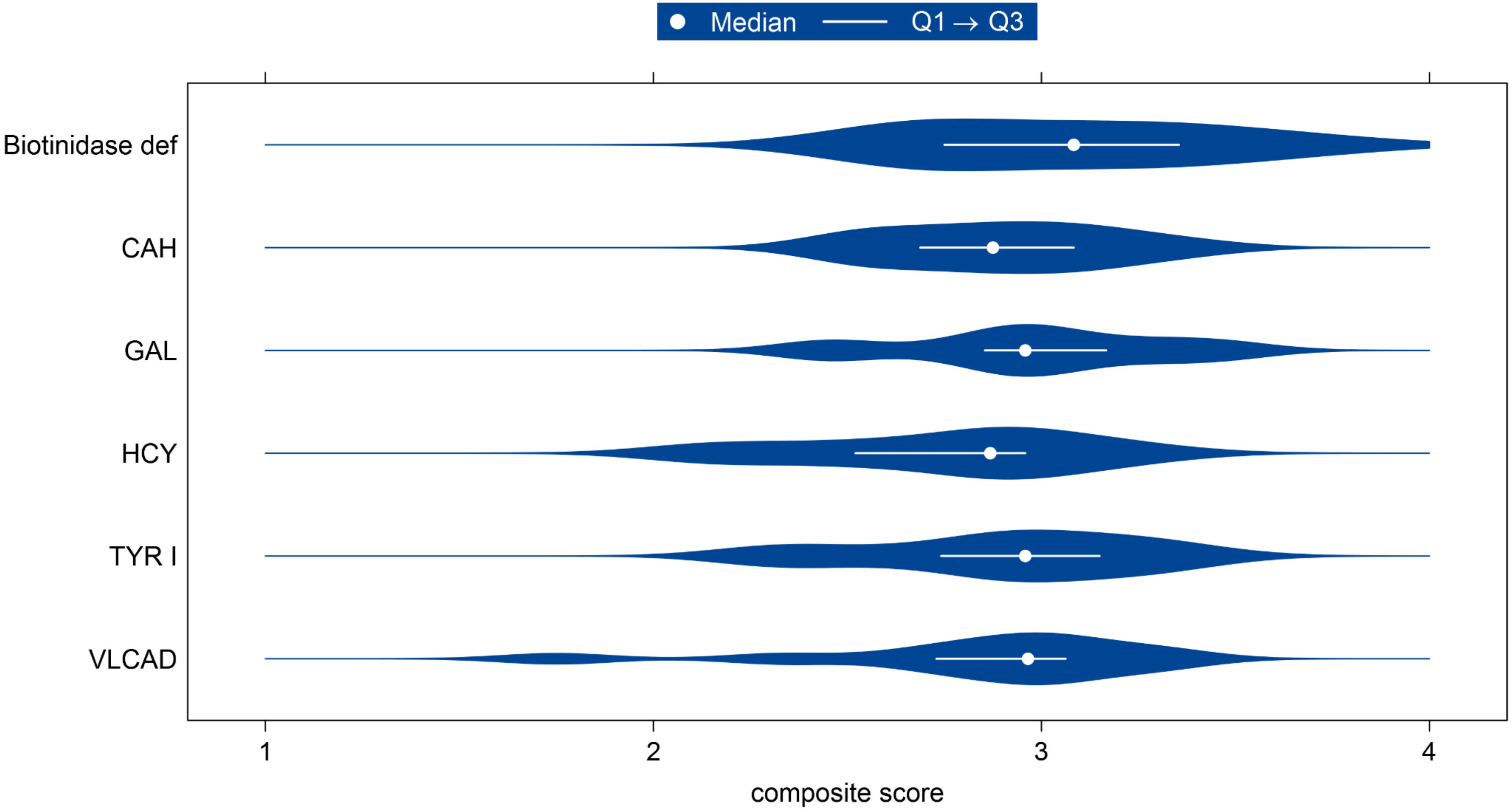
Figure 19 – Ranking of diseases per evaluator (main analysis, n=11)



Number of ranks by evaluator ranking the disease as first (rank 1), second (rank 2) etc.

We show the distribution of the individual composite scores for each of the six diseases by evaluator in the violin plots in Figure 20.

Figure 20 – Distribution of individual composite scores by evaluator (all evaluators)



4.5.2 Stratified analysis of composite scores: labs vs. other

To detect whether the background of the evaluator providing weights for criteria and scores for specific diseases for those criteria is important we performed a limited stratified analysis. Since we only had eleven evaluators this stratified analysis has limitations and was limited to only two categories: laboratory experts performing the NBS analyses (n=5) and other evaluators (decisions makers, academics and patient support groups) (n=6). Results are shown in Figure 21.

No major differences are seen in this analysis on limited numbers. The stratified distribution of the individual composite scores (5 for the labs and 6 for non-laboratory evaluators) is shown in the violin plots in Figure 22 and Figure 23.

Figure 21 – Global composite scores (and range of individual composite scores by evaluator) for labs vs. other evaluators

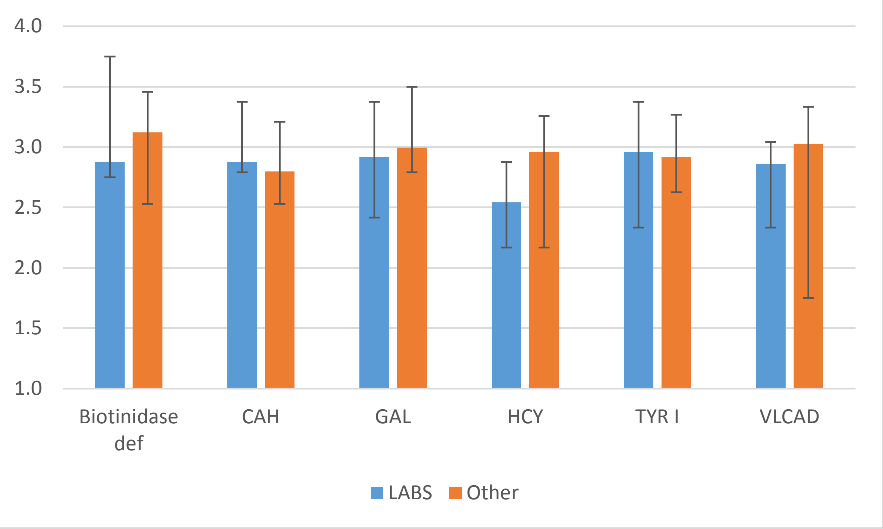


Figure 22 – Distribution of individual composite scores by laboratory evaluators (n=5)

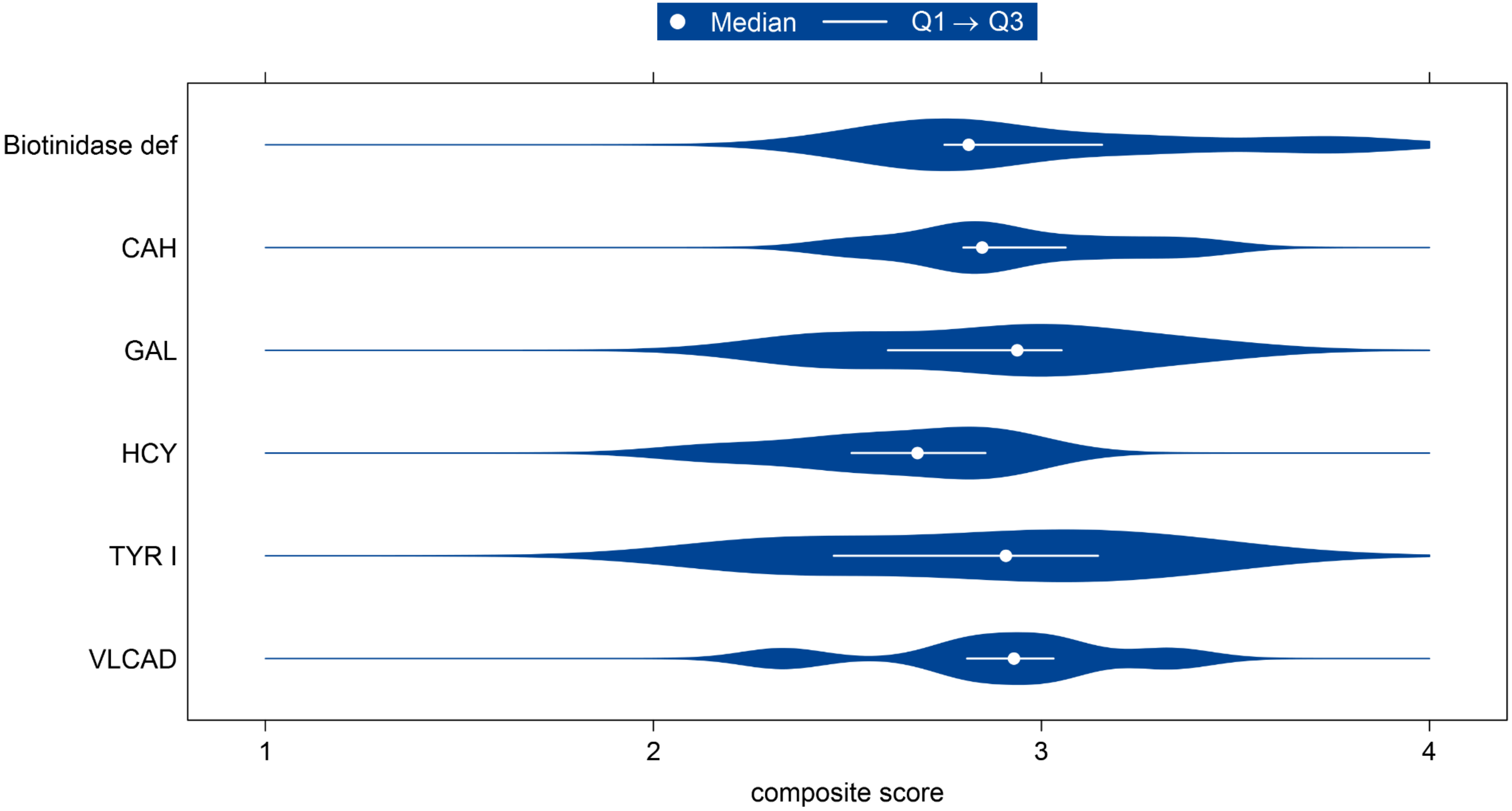
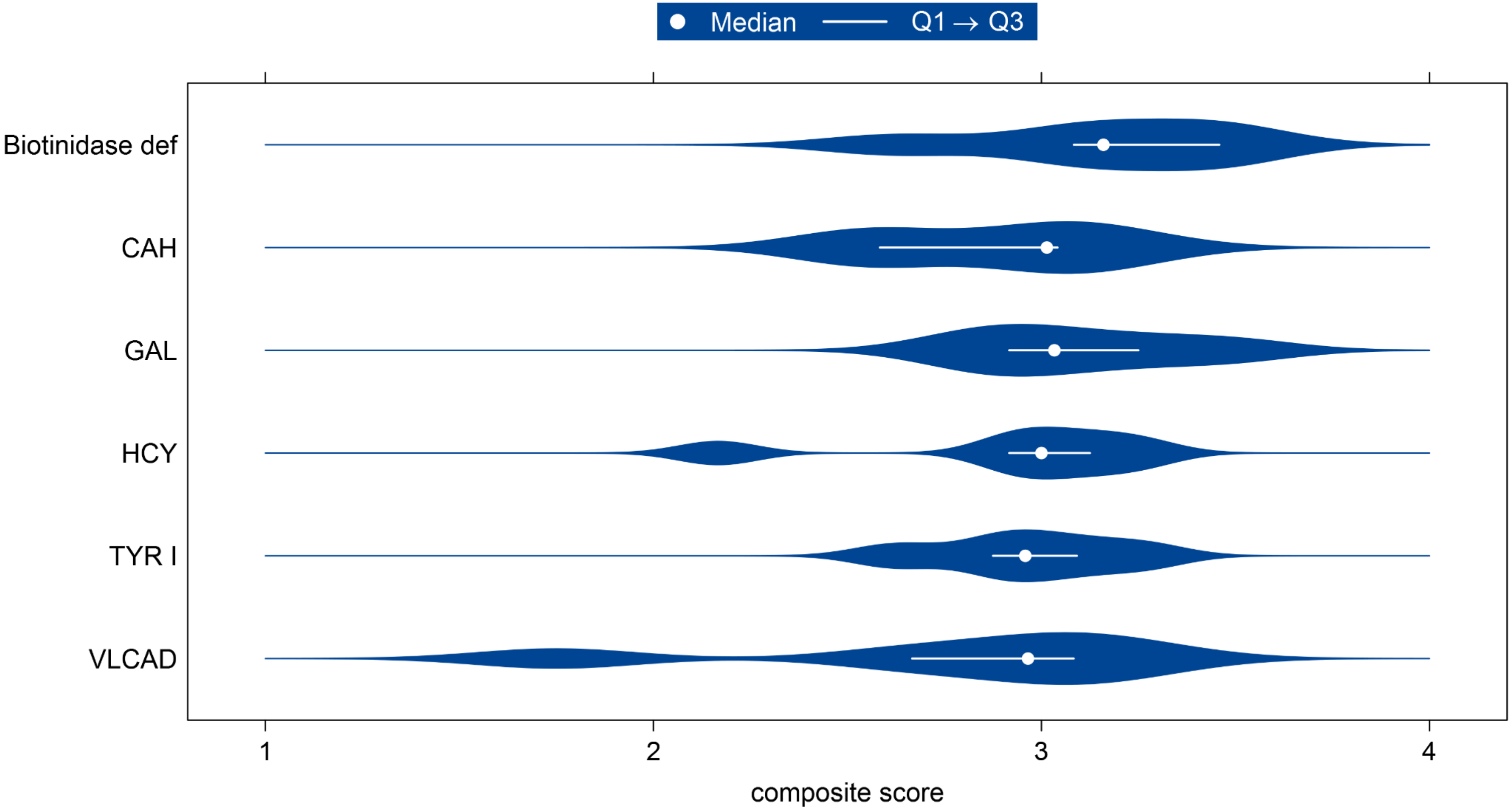


Figure 23 – Distribution of individual composite scores by non-laboratory evaluators (n=6)



4.5.3 Stratified analysis of composite scores: Dutch speaking vs. French speaking community

To detect whether there are differences for the global composite scores for specific diseases by linguistic group we also performed a stratified analysis by community background. Results are shown in Figure 24. One evaluator works for both communities and stated to have given ‘linguistic neutral’ scores. Further distributions of stratified individual scores are shown in Figure 25 and Figure 26.

No major differences are seen in this analysis on limited numbers. Biotidinase deficiency tends to score higher in VG where it is implemented in the NBS. Interestingly, HCY tends to score lower in FWB where it is implemented, compared to VG; however, small numbers of evaluators prevent to establish clear differences. The other diseases show very limited differences in global composite scores between the two groups.

Figure 24 – Global composite scores (and range of individual composite scores by evaluator) for VG and FWB

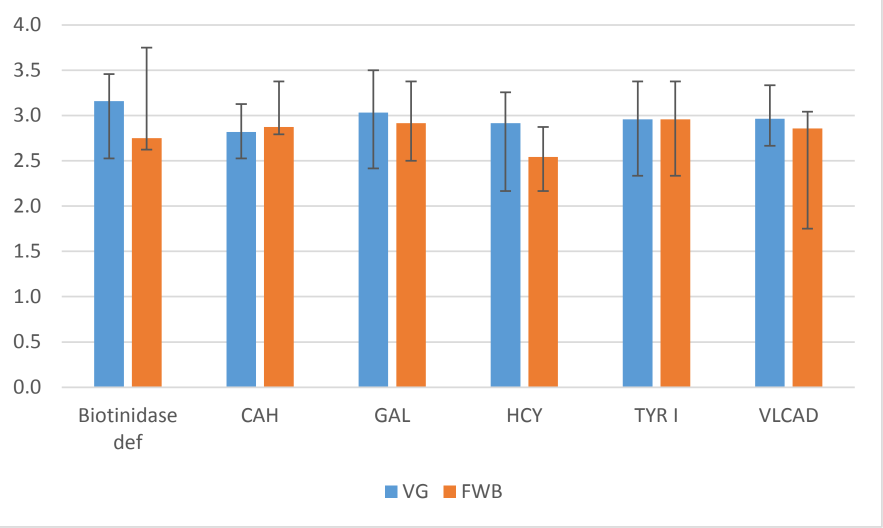


Figure 25 – Distribution of individual composite scores by Dutch speaking community evaluators (n=7)

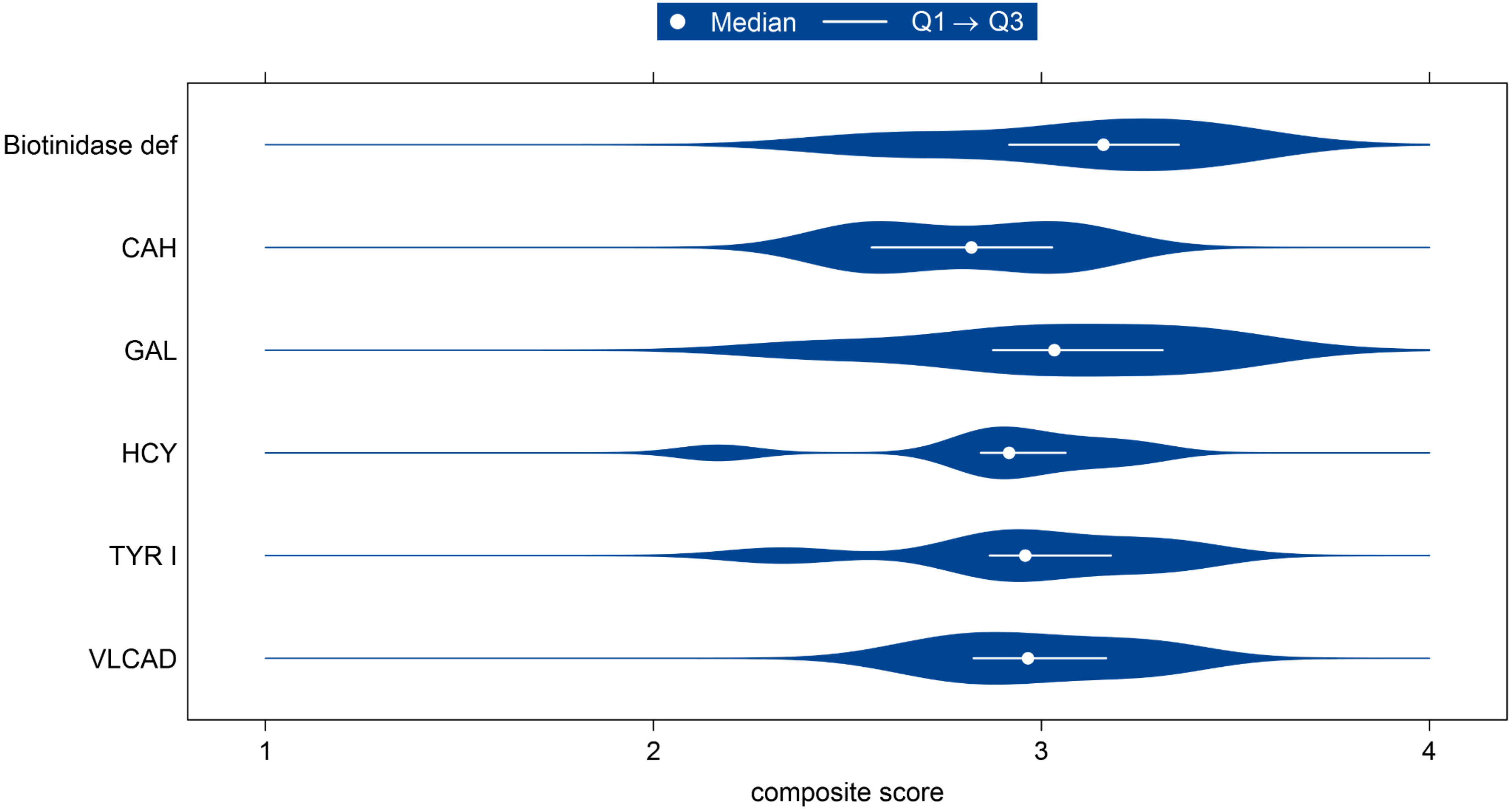
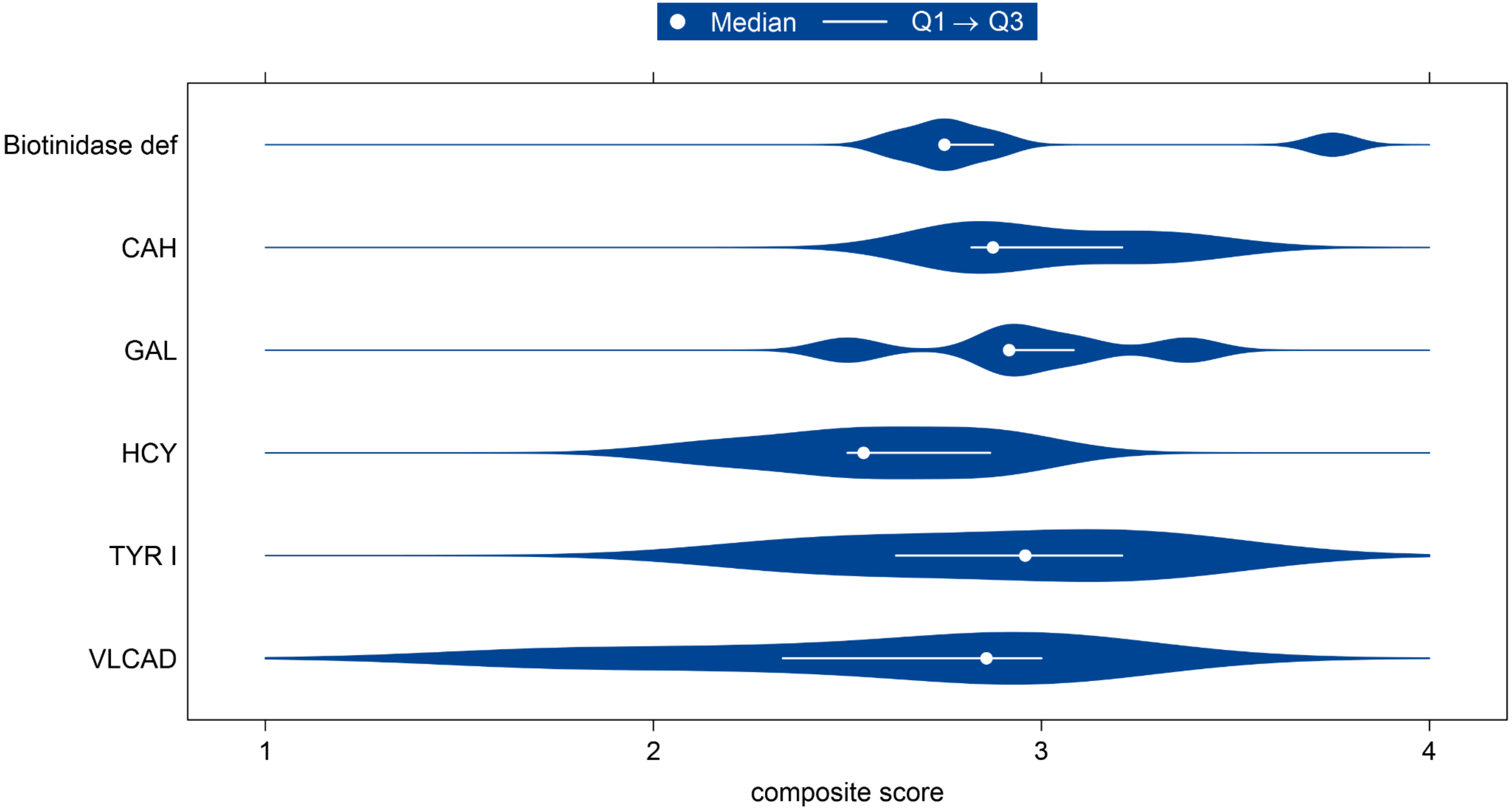


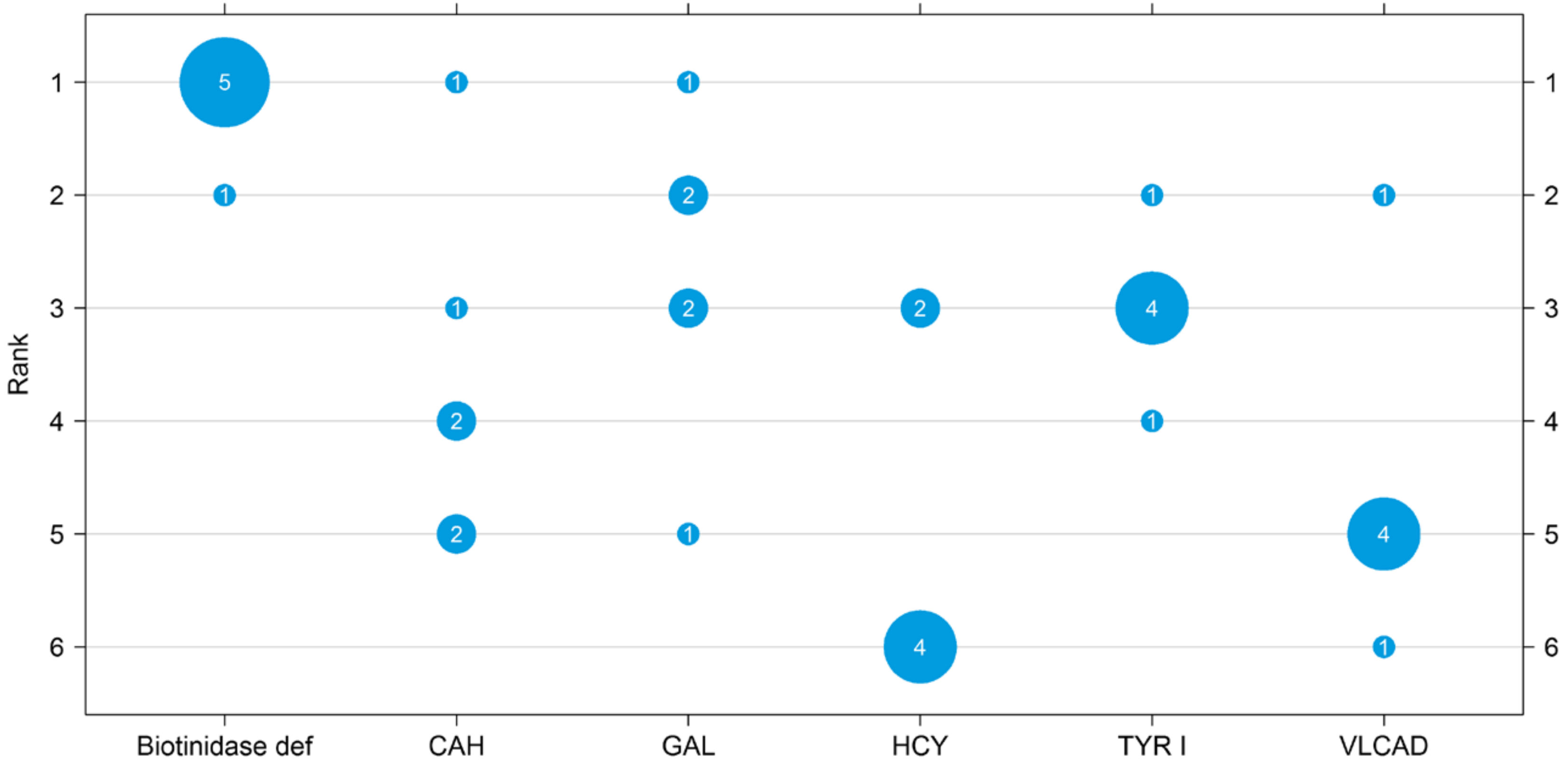
Figure 26 – Distribution of individual composite scores by French speaking community evaluators (n=5)



4.5.4 Global composite scores: additional analyses

As indicated in the methods section we also use five additional methods for calculating global composite scores, for instance by using common scores for all evaluators or by using means. Figure 27 shows that the ranking of each disease is somewhat dependent upon the method chosen to calculate the composite scores. However, we notice some interesting patterns, since biotinidase deficiency is ranked first or second in all methods, GAL is first or second in four of the six methods while HCY and VLCAD is systematically ranked low (because of ex aequo's in the scores the total number by rank is not always six).

Figure 27 – Ranking by disease for the six methods used to calculate composite scores



5 DISCUSSION

We tested this methodology on six pilot diseases. The diseases were selected on the base of one criterion: to be included in the disease list of one community and not in the list of the other community. The diseases included are listed in Table 5. The application of the methodology and the final results allowed to elaborate a list of lessons learned, that were discussed with the steering group participants, and are described below.

Table 5 – Diseases included in this pilot evaluation

Disease	Abbreviation	In Vlaamse Gemeenschap	In Fédération Wallonie-Bruxelles
Congenital Adrenal Hyperplasia	CAH	Yes	No
Biotinidase deficiency	LMCD	Yes	No
Galactosaemias	GAL	No	Yes
Tyrosinemia Type I	TYR I	No	Yes
Homocystinuria	HCY	No	Yes
Very long Chain CoA deshydrogenase deficency	VLCAD	No	Yes

5.1 Selection and definition of criteria

A limited number of seven criteria was selected for the disease scoring, in short: disease frequency, severity of disease in untreated cases, timely availability of the test results, efficacy of early treatment vs. late treatment, probability and impact of false positive results, probability and impact of false negative results, impact on the health care system. The selection and definition of these criteria was intensely discussed with the steering group. A main challenge is that different lists of criteria have been included in the legal texts of each community, and adding all criteria would lead to a long list and a risk of redundancy across criteria. If the same aspects of disease screening are included in more than one criterion, a higher importance will be given to that aspect, leading to imbalanced results in a global score by disease.

Although experts from the steering group proposed to add a number of criteria, it was initially agreed to keep the limited list of seven criteria inspired by the INESSS methodology, adapted and refined in a steering group meeting. However, the supplementary aspects were added into one of the existing criteria, and each criterion was defined into details. In particular, the question of whether to add a specific criterion about the ethics of screening was raised. However, the group agreed that it would be difficult to define which of ethical aspects should be included and in particular to define how they should be scored, i.e. what should be considered as being in favour or in disfavour of disease screening. It was thus proposed to add ethical aspects to each of the seven criteria. A limitation was that no precise definition of the ethical aspects to consider under each criterion was provided. For instance an aspect to consider the probability of false positive results is the (unnecessary) emotional burden on families. It is thus unclear to which extent each participant considered the ethical aspects in its scoring.

The full scoring exercise however allowed to learn lessons on the selected criteria. Despite the definition provided for each criterion, some misunderstanding remained on the precise content of criteria. Some aspects were not - or not sufficiently - clarified in the definitions, as said above for ethical aspects or for the use of positive predictive values in the criteria about false positive results. Likewise, the performance of confirmatory tests were not included in one of the criteria while it may have a large influence on the usefulness of the screening test. Members of the steering group proposed that a more precise description (e.g. checklist format) of the content of each criterion should be provided, a methodological manual should be elaborated and/or in depth discussion between the members of any selection committee using this method should take place to clarify the content of each criterion.

The use of a limited list of criteria, and in particular when criteria contain many different aspects such as impact on the health care system, may increase the risk of misunderstandings about which details should be considered under which criteria. Furthermore, the different aspects included in some complex criteria may go in diverging directions for the scores that would be given to diseases.

5.2 Defining weights of criteria

The weight of each criterion was obtained by asking the steering group members to score them for importance using a four point LIKERT scale. Preliminary results were submitted to a steering group meeting in which some of the criteria were clarified and further defined, and a reweighing was performed after a second meeting.

The means and medians of the weight given to each criterion by the participants do not show substantial variations across criteria. However, a trend for higher weight is observed for the criteria disease severity, efficacy of early vs. late treatment and the probability and impact of false negative results (missed cases). Likewise, no major difference was observed between the median and the mean weight of each criterion but the range of weight by criteria is wider when using the mean (2.4 to 4.0) than when using the median (3 to 4). This pilot analysis did not include any absolute criteria, i.e. a criterion that would imply an “always do” action, or conversely.

Despite this relative homogeneity in central tendency (mean and median), the individual weights given by each participant to each of the criteria show very different distributions across criteria, as could be expected from a group composed of different profiles. The most coherent distribution is observed for disease severity, having received the maximum score by all participants. Consistent unimodal distributions of the weights are seen for efficacy of early vs. late treatment, the probability and impact of false positive results and the impact on the health care system. Higher dispersion across individual weights (bimodal or flat distribution) is observed for disease frequency and probability and impact of false negative results. This suggests very different appreciation between scoring participants of the weight given to these criteria. A number of factors have been discussed in the steering group to explain this discordance in the evaluation of the weight importance:

- The precise definition and correct understanding of each criterion has an impact on the weight given. For instance the timely availability of test results was misunderstood by some experts by being a “short time”, while it is defined as “Is the test result available at a timely moment to prevent preventable complications and sequelae before a diagnosis based on suggestive clinical signs would have been made”. As said above, such analysis would benefit from a more precise description of the content of each criterion.
- The profile of the participants likely influences the weight given to each criterion since the weight given to a criterion may differ according to the viewpoint taken, i.e. varies between public health authority, reference laboratory, curative sector or societal groups, making these weights less universally applicable across an evaluation committee. The personal experience and working methods of some participants could also influence the weight given. From the limited stratified analysis on eleven evaluators it appears that the median weights for the criteria ‘frequency’ and ‘timely availability of test results’ are lower for the lab evaluators but that the criterion ‘probability and impact of false positive results’ is scored somewhat higher by them. For instance the criterion “timely availability of test results” had a low importance for some laboratory experts because they automatically adapt the speed of sending the results for each specific disease (telephone rather than regular mail). The disease frequency did not seem important for some laboratory experts if this represents an additional result of the laboratory technique already used (here MS/MS). This criterion was also less important for other participants because all these diseases are by nature rare. An option would have been to have separate weights between public health decision-makers (community representatives) and disease and screening experts. But this would not make decision-maker easier. Stratified analysis of the criteria weights by type of participant would allow to explore the impact of profile on the final weights. If it is confirmed that the composition of the steering group has an influence on the final outcome (selection of priority diseases).Future prioritisation exercise should pay attention to properly balance the participant profiles.
- The criterion “impact on the health care system” has a large meaning, involving economic and organisational, feasibility and acceptability issues, which makes it difficult to weigh. Moreover, this impact is related to both the screening and the management of detected cases. This mix of levels of health care in the same criterion was debated among participant evaluators because decision making in Belgium differs between the two levels of health care: screening, being part of health

prevention, is decided at the level of the Communities, while case management, being part of curative health care, is funded by the Social security (INAMI – RIZIV) and currently decided at federal level. One option to facilitate decision making would be to separate this criterion into “impact and cost of screening” and “impact on curative health care” involving savings in terms of prevented treatment. However, this would prevent to consider the overall impact of screening for a new disease on the population as a whole.

- Some participants suggested that the weight given to a criterion should vary by disease, e.g. that the probability of a positive test would be important for some diseases and not for others. However, such a differential approach would rely on evaluator judgment, may hamper transparency and would therefore not allow to properly prioritize diseases above others in a comparable manner.

5.3 Scoring diseases

The information on each criterion for each disease was provided to evaluators into a disease summary to provide a homogeneous and more objective basis for scoring. Diseases were also scored on a four point LIKERT scale with an additional zero option when they considered that not enough data were available for scoring (zero values were corrected for in the global composite score). After a preliminary scoring by each participant, a meeting was organised to test the face validity of the results. Criteria and disease information provided were further clarified and participants could adapt their score when there appeared to be any misunderstandings.

One explanation is that for some criteria the data are often incomplete since we are dealing with rare diseases. The personal knowledge and experience of evaluators had likely an important influence on the scores in the absence of sufficiently robust data.

Results of individual scores by disease and criteria show that differences in scores are more often observed for specific criteria such as the probability and impact of false negative results and the impact on the health care system. One explanation suggested by evaluators was that the probability of a false negative result and its impact may have different or even opposite magnitudes for specific diseases (e.g. high probability but low impact). Experts suggested that it would be preferable to separate these two components (probability and impact of test results) into separate criteria. Likewise, the impact of false positive cases may be partly redundant with the impact on the health care system, in particular regarding case confirmation and management. Another factor for divergent answers is the lack of availability of sound data on several criteria as we deal with rare diseases.

The criteria “impact on the health care system” systematically resulted in dispersed results across participants, showing a “flat” distribution for each disease, with scores ranging between 1 to 4 or 2 to 4, while central tendency measures (mean and median) do not reflect these divergences i.e. being mostly around 3. Scoring participants considered that this may be due to the large definition of this complex criterion (as said above), which contains many different components, and to the variation in the profiles of the evaluators. For instance evaluators involved in decision making (e.g. community representatives) can be more sensitive to economic and budget issues, while this may not be considered as important by laboratory experts, who could give more importance to the performance of the screening itself. One option would be to separate this criteria into two components: one economic criterion including cost of screening and cost of case management, cost-effectiveness analyses; and one organisational criterion including continuity of care, feasibility and acceptability. Likewise, community decision makers may be more influenced by the preventive impact while clinicians may be more sensitive to the impact on health care management, as explained above.

This pilot analysis could help detecting possible redundancy between the selected criteria. In particular, the impact of false negative results is likely overlapping with the severity of untreated disease, as the impact of missed cases is considered in terms of severity of the evolution of undiagnosed and thus untreated cases. A method to explore this possible redundancy would be to conduct correlation analyses on the individual scores that are suspected as being redundant, but the steering group was too limited to generate sound conclusions here. If redundancy is confirmed, criteria should be revised.

Due to the variation in the profile of the scoring participants, the stratification of results by type of evaluator could help understand discrepancies in scores. For instance it would be expected that evaluators without specific expertise in the specific disease will likely provide a different weight and there scores will be more based on the information provided in the scientific disease material. In the stratified analyses of the composite scores no major differences were detected neither by function nor by language group. However, this stratified analysis is done on very small numbers so it is not very reliable. However, the composition of the scoring group may have a high influence on the final ranking and should be considered carefully in similar analysis.

5.4 Composite scores by disease and disease prioritization

The most striking finding of this MCDA pilot testing is that total composite scores by disease show little differences across the six selected diseases and thus provide a low discriminative power to select priority diseases among those six specific diseases. As a result, it is difficult to use this ranking for decision making on a screen / do not screen strategy for these diseases. This analysis might be more useful to prioritise a list of diseases against each other, to be possibly used as one of the considerations to decide on screening for a wider list of more different diseases. Furthermore, the ranking of diseases by score changed somewhat according to the method used to calculate the composite score, suggesting that further research into this MCDA method for this type of diseases is needed.

The reasons for these small differences in the global composite scores have been discussed in the steering group. First, the six diseases selected for this pilot testing have probably a similar importance and interest for NBS screening because they are already screened in one part of Belgium, thus were previously considered as being worthwhile to be screened for by a group of decision makers. It would be valuable to repeat this exercise with more extreme diseases, for instance with diseases that are much more likely to rank low or high. Second, the narrow scale for scores, i.e. from 1 to 4, may tend to decrease the potential differences in individual scores. A scale from 1 to 10 might provide different results, although this is not proven in this analysis. Third, the indicators used were central tendencies (median or mean) while the distributions by criteria and disease often showed a high level of dispersion and even divergence, with sometimes bimodal shape at the two extremes of the scale. Correlation and measures of covariance by disease could be more informative than the central tendency alone.

Several experts considered that basing decision-making on such complex issues on a median or sum of weighed scores might not be appropriate. This ranking exercise was not felt sufficient as sole basis for decision-making on screening or not screening for a specific disease. For instance, it may dilute the importance of critical criteria, e.g. a more frequent and very severe disease, easily treatable, with timely results of the test not rank very high because of an expensive test, expensive treatment and a suboptimal test. However, because this scoring induces more reflection and makes arguments explicit, results of this ranking exercise were felt useful to feed the decision-making, together with other arguments.

It should be noted that setting threshold for composite scores to determine whether a disease should be included or excluded from the NBS list of diseases is not the purpose of this study. Such decisional tools can only be developed at the decision-maker level, i.e. the communities. Moreover, the technology for screening and the context change rapidly, therefore it might be important to repeat this exercise regularly since the evaluation can differ depending on future changes.

5.5 Similar experience in other countries

The INESSS evaluation in Quebec allowed to rank diseases by priority order for inclusion in the NBS programme, and the ranking was used to define three separate waves of disease that will be introduced in the NBS programme. The INESSS analysis also showed limited differences in composite score across diseases but this aspect is not further commented in the report.

5.6 Conclusions

In this study, we test an MCDA method to rank diseases that can be detected through neonatal blood screening by priority. Building a consensus about the selection and uniform definitions of the criteria proved to be more complex than anticipated, and required a considerable amount of time and exchanges. In this study, global composite scores show little differences across the six selected diseases and thus provide a low discriminative power to rank diseases for NBS.

The ranking of diseases changed somewhat according to the method used to calculate the composite scores, making further discussion about this method necessary.

Several members of the steering group considered it inappropriate to base policy and decision-making on such complex issue only on a composite score by disease. However, because this weighing and scoring induces more reflection and prompts to make arguments explicit, results of this exercise were felt useful to feed the decision-making.

This method could be improved by developing a more detailed definition of each criterion, splitting up criteria that may have antagonist components, considering a larger scale to weigh and score, and by paying careful attention to the balance of the profiles of those participating in the evaluation.

Future exercises to rank diseases should preferably include a wider range of diseases and can benefit from the lessons learned in this pilot test.

5.7 Key messages and lessons learned for the future

- Decision making about including an (additional) disease in a neonatal blood screening programme is a complex matter with many dimensions. For those decision making processes formal decision-making methods are increasingly used.
- Inspired by the experience of the INESSS agency in Québec we conducted a pilot study to test the usability of a *Multi-Criteria Decision Analysis (MCDA)*. Together with a steering group with representatives and experts of the communities in Belgium and stakeholders we decided to test this instrument on the six diseases that are currently only screened for in one of the two communities.
- The selection of the decision criteria is not easy. The heterogeneous weights given for some criteria show that this process should preferably run through multiple cycles to specify the precise content of the criteria and, if necessary, to add content or to split criteria with the aim to avoid misunderstandings.
- Evaluators can have diverging points-of-view. Attributing a weight and score for each of the criteria calls for a broad discussion between evaluators.
- Scoring diseases for each of those criteria makes it unavoidable to collect information for each disease and for each of those criteria; This information should include evidence on epidemiology, available tests and interventions, including the organisational, health economical and ethical aspects. This information should be made available to all evaluators so they can all judge with the same basic information.
- To determine the weights for each criterion and for the scoring of the diseases for each of those criteria a LIKERT scale with only four points may be too limited and not enough discriminating. We could not really answer this question. The six diseases were, for decision making, probably in the same zone between acceptable or non-acceptable for a screening programme, in the first place since they had already been selected by one community.
- During the composition of the panel of evaluators careful attention should be paid to a balance between them to make sure that they represent all relative viewpoints and to avoid that essential arguments are overlooked.
- During the steering board meetings it was mentioned that it might be useful to consider an ‘exclusion score’ to indicated if a necessary condition for screening a specific disease is not fulfilled.
- The exact way to calculate global composite scores slightly influences the global composite score and the ranking of diseases. Again, with the six selected diseases we could not really answer this question as the global composite scores were too close.
- It is not the aim to determine decision making on inclusion or exclusion of specific disease in a screening programme exclusively on a MCDA ranking. The aim of this exercise is to allow discussions about this decision making to be more objective and transparent. This way, a better cohesion between successive discussions can be reached.

■ REFERENCES

The references to the scientific information pertaining to the six specific diseases can be found in the supplement.

1. Burgard P, Cornel M, Di Filippo F, Haege G, Hoffmann GF, Lindner M, et al. Report on the practices of newborn screening for rare disorders implemented in Member States of the European Union, Candidate, Potential Candidate and EFTA Countries. 2012. EU network of experts on Newborn Screening Available from: http://ec.europa.eu/health/rare_diseases/screening/index_en.htm
2. Koninklijk besluit betreffende de erkenning van de diensten voor opsporing van de fenylcetonurie en de toekenning van subsidies aan die diensten. 1968.
3. Arrêté royal relatif à l'agréation des services de dépistage des anomalies congénitales métaboliques et à l'octroi de subventions à ces services, 1974.
4. Goyens P, Laeremans H. Le dépistage néonatal en Fédération Wallonie - Bruxelles. InfONE. 2014(2):6.
5. 27 MAI 2009. Arrêté du Gouvernement de la Communauté française en matière de dépistage des anomalies congénitales en Communauté française, Moniteur Belge 2009.
6. Toussaint B, Pereira T, Goyens P, Laeremans H, Vincent M, Marie S, et al. Guide pour le programme de dépistage néonatal des anomalies métaboliques en FWB. In; 2013.
7. 22 MAI 2014 - Arrêté du Gouvernement de la Communauté française fixant le protocole du programme de dépistage des anomalies congénitales en Communauté française, 2014.
8. Van Rossom J. Maladies métaboliques: dépister pour mieux soigner. Osiris News. 2012(26).
9. Dépistage néonatal des maladies métaboliques et endocriniennes [Web page].2014. Available from: http://www.saintluc.be/laboratoires/activites/biologie-clinique/biochimie/depistage-neonatal.php
10. Vlaams Agentschap Zorg en Gezondheid. Vlaams bevolkingsonderzoek naar aangeboren aandoeningen bij pasgeborenen via een bloedstaal: Draaiboek 2012. 2012. Available from: http://www.zorg-en-gezondheid.be/uploadedFiles/Zorg_en_Gezondheid/Ziektes/Aangeboren_aandoeningen/AAP_draaiboek_2012.pdf

11. Ministerieel Besluit ter aanvulling van het ministerieel besluit van 18 maart 1974, houdende uitvoering van het koninklijk besluit van 13 maart 1974 betreffende de erkenning van de diensten voor opsporing van de aangeboren metabolische afwijkingen en de toekenning van subsidies aan die diensten, gewijzigd bij het koninklijk besluit van 25 april 1980. , 1980.
12. Besluit van de Vlaamse regering betreffende de centra voor opsporing van de aangeboren metabolische afwijkingen., 1997.
13. Ministerieel besluit tot regeling van de werkings- en erkenningsprocedure betreffende de centra voor de opsporing van aangeboren metabolische afwijkingen., 1998.
14. Decreet betreffende het preventieve gezondheidsbeleid., 2003.
15. Besluit van de Vlaamse Regering betreffende bevolkingsonderzoek in het kader van ziektepreventie., 2008.
16. Vlaams Agentschap Zorg en Gezondheid. Vlaams bevolkingsonderzoek naar aangeboren aandoeningen bij pasgeborenen via een bloedstaal: Draaiboek 2015. 2015.
17. Ministerieel besluit tot oprichting van de Vlaamse werkgroep Bevolkingsonderzoek naar aangeboren aandoeningen bij pasgeborenen via een bloedstaal, 2012.
18. Côté B, Gosselin C, Renaud J. Pertinence d'élargir le programme de dépistage néonatal sanguin au Québec. Québec: INESSS; 2013. ETMIS 2013; Vol 9 N° 7 Available from: https://www.inesss.qc.ca/publications/publications/publication/pertinence-delargir-le-programme-de-depistage-neonatal-sanguin-au-quebec.html
19. Health Council of the Netherlands. Neonatal screening; new recommendations. The Hague: Health Coucil of the Netherlands; 2015. 2015/08 Available from: http://www.gezondheidsraad.nl/sites/default/files/201508neonatale_screening.pdf
20. Proesmans M, Cuppens H, Vincent M-F, Palem A, De Boeck K, Dierickx K, et al. Is neonatal screening for cystic fibrosis recommended in Belgium? Health Technology Assessment (HTA). Brussels: Belgian Health Care Knowledge Centre (KCE); 2010 15/07/2010. KCE Reports 132 (D/2010/10.273/43) Available from: https://kce.fgov.be/publication/report/is-neonatal-screening-for-cystic-fibrosis-recommended-in-belgium
21. The Evidem framework. Available from: https://www.evidem.org/
22. Wilson JMG, Jungner G. Principles and practive of screening for disease. Geneva: World Health Organisation; 1968. Available from: http://apps.who.int/iris/handle/10665/37650
23. Mason CA, Kirby RS, Sever LE, Langlois PH. Prevalence is the preferred measure of frequency of birth defects. Birth Defects Res A Clin Mol Teratol. 2005;73(10):690-2.
24. Department for Communities and Local Government: London. Multi-criteria analysis: a manual. 2009. Available from: www.communities.gov.uk
25. Cleemput I, Devriese S, Kohn L, Devos C, van Til J, Groothuis-Oudshoorn K, et al. Incorporating societal preferences in reimbursement decisions – Relative importance of decision criteria according to Belgian citizens. Health Services Research (HSR) Brussels: Belgian Health Care Knowledge Centre (KCE). 2014. KCE Reports Report 234 Available from: https://kce.fgov.be/publication/report/incorporating-societal-preferences-in-reimbursement-decisions-%E2%80%93-relative-importan
26. Statistics Belgium. 2015. Available from: http://statbel.fgov.be/nl/statistieken/opendata/datasets/bevolking/geboorten/
27. Orphanet. Prevalence and incidence of rare diseases. Bibliographic data. Diseases listed by decreasing prevalence, incidence or number of published cases. Paris: 2015. Orphanet Report Series Available from: http://www.orpha.net/orphacom/cahiers/docs/GB/Prevalence_of_rare_diseases_by_decreasing_prevalence_or_cases.pdf
28. Newborn screening Ontario. 2015. Available from: http://www.newbornscreening.on.ca/bins/content_page.asp?cid=7-21

29. The Online Metabolic and Molecular Bases of Inherited Disease. Available from: http://ommbid.mhmedical.com/

30. National Institutes of Health. Genetics Home Reference. Available from: http://ghr.nlm.nih.gov/