ORIGINAL RESEARCH ARTICLE | Genetics inMedicine

Enhanced interpretation of newborn screening results without analyte cutoff values

Gregg Marquardt, MSS[1], Robert Currier, PhD[2], David M.S. McHugh[1], Dimitar Gavrilov, MD, PhD[1], Mark J. Magera[1], Dietrich Matern, MD[1], Devin Oglesbee, PhD[1], Kimiyo Raymond, MD[1], Piero Rinaldo, MD, PhD[1], Emily H. Smith, PhD[1], Silvia Tortorelli, MD, PhD[1], Coleman T. Turgeon[1], Fred Lorey, PhD[2], Bridget Wilcken, MD[3], Veronica Wiley, PhD[3], Lawrence C. Greed, BSc[4], Barry Lewis, MD[4], François Boemer, PharmD, PhD[5], Roland Schoos, PhD[5], Sandrine Marie, PhD[6], Marie-Françoise Vincent, MD, PhD[6], Yuri Cleverthon Sica, Msc[7], Mouseline Torquado Domingos[7], Khalid Al-Thihli, MD[8], Graham Sinclair, PhD[8], Osama Y. Al-Dirbashi, PhD[9], Pranesh Chakraborty, MD[9], Mark Dymerski[10], Cory Porter[10], Adrienne Manning[11], Margretta R. Seashore, MD[12], Jonessy Quesada, MD[13], Alejandra Reuben[13], Petr Chrastina, MSc[14], Petr Hornik, PhD[14], Iman Atef Mandour, MD[15], Sahar Abdel Atty Sharaf, MD[15], Olaf Bodamer, MD, PhD[16], Bonifacio Dy, MD[17], Jasmin Torres[17], Roberto Zori, MD[18], David Cheillan, PhD[19], Christine Vianey-Saban, PhD[19], David Ludvigson[20], Adrya Stembridge[21], Jim Bonham, PhD[22], Melanie Downing, Msc[22], Yannis Dotsikas, PhD[23], Yannis L. Loukas, PhD[23], Vagelis Papakonstantinou, PhD[24], Georgios S.A. Zacharioudakis, PhD[24], Ákos Baráth, PhD[25], Eszter Karg, MD, PhD[25], Leifur Franzson, PhD[26], Jon J. Jonsson, MD, PhD[26], Nancy N. Breen[27], Barbara G. Lesko[27], Stanton L. Berberich, PhD[28], Kimberley Turner, RN[29], Margherita Ruoppolo, MD[30], Emanuela Scolamiero[30], Italo Antonozzi, MD[31], Claudia Carducci, MS[31], Ubaldo Caruso[32], Michela Cassanello[32], Giancarlo la Marca, Pharm Sc[33], Elisabetta Pasquini, MD[34], Iole Maria Di Gangi, PhD[35], Giuseppe Giordano, PhD[35], Marta Camilot, PhD[36], Francesca Teofoli[36], Shawn M. Manos, BS[37], Colleen K. Peterson, BS[37], Stephanie K. Mayfield Gibson, MD[38], Darrin W. Sevier[38], Soo-Youn Lee, MD, PhD[39], Hyung-Doo Park, MD, PhD[39], Issam Khneisser, MS[40], Phaidra Browning[41], Fizza Gulamali-Majid, PhD[42], Michael S. Watson, PhD[43], Roger B. Eaton, PhD[44], Inderneel Sahai, MD[44], Consuelo Ruiz[45], Rosario Torres[45], Mary A. Seeterlin, PhD[46], Eleanor L. Stanley[46], Amy Hietala[47], Mark McCann[47], Carlene Campbell[48], Patrick V. Hopkins[48], Monique G. de Sain–Van der Velden, PhD[49], Bert Elvers[50], Mark A. Morrissey, PhD[51], Sherlykutty Sunny[51], Detlef Knoll, MSc[52], Dianne Webster, PhD[52], Dianne M. Frazier, PhD[53], Julie D. McClure, MPH[53], David E. Sesser[54], Sharon A. Willis[54], Hugo Rocha, Msc[55], Laura Vilarinho, PhD[55], Catharine John, PhD[56], James Lim, PhD[56], S. Graham Caldwell[57], Kathy Tomashitis, MNS[57], Daisy E. Castiñeiras Ramos[58], Jose Angel Cocho de Juan, PhD[58], Inmaculada Rueda Fernández, MD[59], Raquel Yahyaoui Macías, MD, PhD[59], José María Egea-Mellado[60], Inmaculada González-Gallego, PhD[60], Carmen Delgado Pecellin, PhD[61], Maria Sierra García-Valdecasas Bermejo, PhD[61], Yin-Hsiu Chien, MD, PhD[62], Wuh-Liang Hwu, MD, PhD[62], Thomas Childs, MT(ASCP)[63], Christine D. McKeever[63], Tijen Tanyalcin, MD, PhD[64], Mahera Abdulrahman, MD, PhD[65], Cecilia Queijo, PhD[66], Aída Lemes, MD[66], Tim Davis[67], William Hoffman[67], Mei Baker, MD[68] and Gary L. Hoffman[68]

[1]Department of Laboratory Medicine and Pathology, Mayo Clinic College of Medicine, Rochester, Minnesota, USA; [2]California Department of Public Heath, Richmond, California, USA; [3]The Children's Hospital at Westmead, Sydney, New South Wales, Australia; [4]Department of Clinical Biochemistry, Princess Margaret Hospital, Perth, Western Australia, Australia; [5]Centre Hospitalier Universitaire de Liège, Liège, Belgium; [6]Cliniques Universitaires Saint-Luc, Université Catholique de Louvain, Brussels, Belgium; [7]Fundação Ecumênica de Proteção ao Excepcional, Curitiba, Brazil; [8]Children's & Women's Health Center, Vancouver, British Columbia, Canada; [9]Children's Hospital of Eastern Ontario, Ottawa, Ontario, Canada; [10]Colorado Department of Public Health and Environment, Denver, Colorado, USA; [11]Connecticut Department of Public Health Laboratory, Hartford, Connecticut, USA; [12]Department of Genetics, Yale University School of Medicine, New Haven, Connecticut, USA; [13]Hospital Nacional de Niños, San José, Costa Rica; [14]Institute of Inherited Metabolic Disorders, First Faculty of Medicine, Charles University and General University Hospital, Prague, Czech Republic; [15]Cairo University Faculty of Medicine, Cairo, Egypt; [16]Department of Human Genetics, University of Miami Miller School of Medicine, Miami, Florida, USA; [17]Florida Newborn Screening Program, Jacksonville, Florida, USA; [18]Department of Pediatrics, University of Florida, Gainesville, Florida, USA; [19]Hospices Civils de Lyon, Lyon, France; [20]Georgia Department of Public Health, Atlanta, Georgia, USA; [21]Department of Human Genetics, Emory University, Atlanta, Georgia, USA; [22]Sheffield Children's NHS Foundation Trust, Sheffield, UK; [23]School of Pharmacy, University of Athens, Athens, Greece;

Submitted 31 August 2011; accepted 28 December 2011; advance online publication 16 February 2012. doi:10.1038/gim.2012.2

(Affiliations continue on the following page)

Purpose: To improve quality of newborn screening by tandem mass spectrometry with a novel approach made possible by the collaboration of 154 laboratories in 49 countries.

Methods: A database of 767,464 results from 12,721 cases affected with 60 conditions was used to build multivariate pattern recognition software that generates tools integrating multiple clinically significant results into a single score. This score is determined by the overlap between normal and disease ranges, penetration within the disease range, differences between conditions, and weighted correction factors.

Results: Ninety tools target either a single condition or the differential diagnosis between multiple conditions. Scores are expressed as the percentile rank among all cases with the same condition and are compared to interpretation guidelines. Retrospective evaluation of past cases suggests that these tools could have avoided at least half of 279 false-positive outcomes caused by carrier status for fatty-acid oxidation disorders and could have prevented 88% of known false-negative events.

Conclusions: Application of this computational approach to raw data is independent from single analyte cutoff values. In Minnesota, the tools have been a major contributing factor to the sustained achievement of a false-positive rate below 0.1% and a positive predictive value above 60%.

Genet Med advance online publication 16 February 2012

Key words: cutoff values; false-positive rate; inborn errors of metabolism; newborn screening; positive predictive value.

INTRODUCTION

The Regional Genetics and Newborn Screening Collaboratives funded by the Maternal and Child Health Bureau have been very successful in improving the newborn screening infrastructure of the United States. One of these initiatives has supported a project to hasten the implementation of newborn screening by tandem mass spectrometry (MS/MS)[1] and achieve uniformity of targets.[2] The importance of this endeavor is underscored by the recent inclusion of expanded newborn screening among the 10 great public health achievements of the past decade in the field of maternal and infant health.[3]

The specific objectives of the collaborative project are (i) to achieve consistency with the uniform panel adopted as the national standard by the Secretary of Health and Human Services[4] and (ii) to improve analytical performance through the pursuit of the lowest achievable rates of false-positive and false-negative results.[5] This project has grown to include 154 public health programs and private laboratories worldwide, leading to the publication of 8,255 disease ranges and 114 cutoff target ranges for amino acids, acylcarnitines, and related ratios.[6,7]

We have developed multivariate pattern-recognition software designed to convert metabolic profiles into a composite score driven by the degree of overlap between normal population and disease range. Clinical relevance of a marker is reached when the median of the disease range is outside the percentile limits of the normal population.[6] A simultaneous assessment of multiple analytes is performed according to the degree of penetration within the respective disease range, expected differences between specific conditions, and proportionally weighted correction factors. This approach could represent a viable alternative to analyte cutoff values in the process of raw data interpretation, fostering their replacement with score-interpretation guidelines for a given condition.

MATERIALS AND METHODS

The Region 4 Stork MS/MS data project is a Web-based application developed using Microsoft.NET framework 3.5 and SQL Server 2008.[6] The criteria for case definition are set by the local protocols of the individual participating sites and by overarching requirements that have been described previously.[6] As of 15 December 2011, the MS/MS profiles of 12,077 patients affected with 60 metabolic disorders and of 644 heterozygote carriers for 12 conditions have been collected in this database. These profiles have served as the training set for the development of the postanalytical tools, and their number continues to expand. Since the beginning of 2009, an average of 5.2 new cases has been added per day (2008: 1,796 cases; 2009: 1,734 cases; 2010: 1,452 cases). The current population study translates to 767,408 discrete analyte concentrations and calculated

[24]Neolab SA, Athens, Greece; [25]Department of Pediatrics, University of Szeged, Szeged, Hungary; [26]Department of Genetics and Molecular Medicine, Landspitali University of Iceland, Reykjavik, Iceland; [27]Indiana Newborn Screening Laboratory, Indianapolis, Indiana, USA; [28]State Hygienic Laboratory at the University of Iowa, Ankeny, Iowa, USA; [29]University of Iowa Children's Hospital, Iowa City, Iowa, USA; [30]CEINGE—Biotecnologie Avanzate, Università degli Studi di Napoli "Federico II," Naples, Italy; [31]Sapienza University of Rome, Rome, Italy; [32]University Department of Pediatrics, G. Gaslini Institute, Genoa, Italy; [33]Department of Pharmacology, University of Florence, Florence, Italy; [34]Meyer Children's Hospital, Florence, Italy; [35]Department of Pediatrics, Università di Padova, Padua, Italy; [36]Azienda Ospedaliera Universitaria Integrata di Verona, Verona, Italy; [37]Kansas Department of Health & Environmental Laboratories, Topeka, Kansas, USA; [38]Kentucky Department for Public Health, Frankfort, Kentucky, USA; [39]Sungkyunkwan University School of Medicine, Seoul, South Korea; [40]Saint Joseph University, Beirut, Lebanon; [41]Tulane University, New Orleans, Louisiana, USA; [42]Department of Health and Mental Hygiene, Baltimore, Maryland, USA; [43]American College of Medical Genetics, Bethesda, Maryland, USA; [44]New England Newborn Screening Program, Boston, Massachusetts, USA; [45]Universidad Autónoma de Nuevo León, Monterrey, Mexico; [46]Michigan Department of Community Health, Lansing, Michigan, USA; [47]Minnesota Department of Health, St. Paul, Minnesota, USA; [48]Missouri Public Health Laboratory, Jefferson City, Missouri, USA; [49]University Medical Center, Utrecht, The Netherlands; [50]National Institute for Public Health and the Environment, Bilthoven, The Netherlands; [51]Wadsworth Center, New York State Department of Health, Albany, New York, USA; [52]LabPlus, Auckland Hospital, Auckland, New Zealand; [53]Division of Genetics and Metabolism, University of North Carolina, Chapel Hill, Chapel Hill, North Carolina, USA; [54]Oregon State Public Health Laboratory, Hillsboro, Oregon, USA; [55]National Institute of Health Doutor Ricardo Jorge, Porto, Portugal; [56]KK Women's and Children's Hospital, Singapore, Singapore; [57]South Carolina Department of Health and Environmental Control, Columbia, South Carolina, USA; [58]Hospital Clínico Universitario, Santiago de Compostela, Spain; [59]Carlos Haya University Hospital, Málaga, Spain; [60]Lab Methabolopathies, Centro de Bioquímica y Genetica Clínica, H.U. Virgen de la Arrixaca, Murcia, Spain; [61]Unidad de Metabolopatias del Hospital Universitario Virgen del Rocio, Seville, Spain; [62]National Taiwan University Hospital, Taipei, Taiwan; [63]Tennessee Department of Health Laboratory Services, Nashville, Tennessee, USA; [64]Tanyalcin Medical Lab Selective Newborn Screening and Metabolism Unit, Izmir, Turkey; [65]Dubai Genetics Centre, Dubai, United Arab Emirates; [66]Instituto de Seguridad Social, Montevideo, Uruguay; [67]Washington State Department of Health, Shoreline, Washington, USA; [68]Wisconsin State Laboratory of Hygiene, University of Wisconsin, Madison, Wisconsin, USA. Correspondence: Piero Rinaldo (rinaldo@mayo.edu)

ratios. Each case is assigned a unique code separate from any other traceable identifier, and no demographic information is collected except the calendar year of birth. Accordingly, this project has been reviewed and approved as a minimum-risk protocol by the Mayo Clinic Institutional Review Board (protocol PR09-001709-01).

The process and criteria used to create a tool are described in the **Supplementary Material** online. Tools can be generated for one or more conditions following a stepwise process that has four major components (**Supplementary Table S1** online): (i) choice of scoring strategy and method to calculate correction factors; (ii) selection of markers; (iii) activation of differentiators, outlier rules, and filters; and (iv) setup of interpretation guidelines. Different scoring strategies are available to elevate scores for conditions that have only a few informative markers (**Supplementary Table S2** online). The correction factors, which can be either condition- or case-specific, are derived from the degree of overlap between the normal population and the disease range of each informative marker in a given condition. The degree of overlap is indeed the foundation of this novel method for interpreting quantitative results in a way that is unique to each condition and therefore not dependent on fixed analyte cutoff values. The selection of markers is based on an objective threshold of clinical significance, which is reached when the median of the disease range of a marker is above the 99 percentile of the normal population (high markers—i.e., abnormal when above the normal range) or below the 1 percentile (low markers).[6] Differentiators, outlier rules, and filters are added to mitigate the potential impact of true negative cases (cases with completely normal results) to preserve the integrity of the tools and allow differential diagnosis between conditions.

As of 15 December 2011, a total of 90 active tools were accessible on the website, 37 of which are applicable to the differential diagnosis of two or more conditions (**Supplementary Figure S1** online). Their intended use is to generate a score that drives the interpretation and resolution of cases with potentially abnormal MS/MS results. Case profiles can be entered individually (i.e., after the conventional flagging of abnormal results according to cutoff values, **Supplementary Figure S2** online) or as batches containing many profiles (e.g., entire plates/daily runs) uploaded to the website using a health information exchange system.[8]

RESULTS

This multivariate pattern-recognition software is applicable to a broad range of clinical applications. Expanded newborn screening is ideal for a clinical validation study because it involves many markers requiring pattern recognition and profile interpretation. Their complexity is compounded by the rarity of most of the target conditions. At this stage, tools are based on data from neonatal blood spots and are not applicable to different specimen types and to older patients.

Figure 1 shows a partial view of the tool (the data-entry window is not shown; see **Supplementary Material** online for an example of that panel) for argininosuccinic acid lyase deficiency,[9] a urea-cycle disorder that is included in the recommended uniform panel.[2] The top part of the figure is a visual overlay of three elements for each informative analyte (red) and discriminator (gray)—the normal population range, the disease range, and the individual value—all shown after conversion to the multiple of the normal median on a log scale. The screening results of this particular case were not considered informative according to the cutoff value for citrulline applied by the testing laboratory at the time. The bottom part of the figure summarizes the calculated score as follows: (i) the absolute value of the calculated score; (ii) the percentile rank of the score in comparison to all available cases; (iii) the number of available cases with the condition under evaluation; and (iv) a visual display of all scores in comparison to interpretation guidelines. These are built as intervals where the score is considered as either being not informative or indicating that the condition is possible, likely, or very likely. Notably, in this false-negative case the score percentile rank was 29% (N = 78) even with the omission from the tool of the unique marker of this condition, argininosuccinic acid. Following this event, the cutoff value of the program was reduced by 25% and this tool is being used on a regular basis. As of 15 December 2011, 110 of the laboratories participating in the collaborative project have implemented a high cutoff value for citrulline and therefore are bound to encounter cases with hypercitrullinemia in disorders besides citrullinemia type I; the most common among these is indeed argininosuccinic acid lyase deficiency. Because 63% of laboratories have a cutoff value above the recommended target range for citrulline (30–40 μmol/liter),[6] they are likely to experience false-negative events like the one shown here. This is not a rare situation. The project database includes 86 cases (0.7% of the total count) that were reported as normal but in which a later diagnosis was based on clinical presentation. This set of cases is limited to those for which all the results required to calculate a score were available, but there are others, some extracted from the literature, with partial sets of data. Excluding conditions in which the poor sensitivity is driven by either a true lack of an informative marker (nonketotic hyperglycinemia) or the historical reliance on an ineffective marker (tyrosinemia type I),[6] 88% of the remaining cases (59 of 67 patients affected with 23 conditions) generated an informative score when evaluated with the pertinent tool. Six of the eight false-negative cases with uninformative scores have been published.[10–12] Overall, this anecdotal evidence suggests that, pending a prospective study of the impact of the interpretive tools, at least half of historical false-negative events could perhaps have been avoided if these tools had been available and utilized.

Although sensitivity is of critical importance, the greatest opportunity for performance improvement in newborn screening, especially in a multiplex test environment, is found in the realm of specificity. The false-positive rate limited to testing by tandem mass spectrometry ranges between 5.99% and 0.03% (median: 0.46%) among the 68 sites that have shared their performance metrics on the project website. A significant issue that drives high false-positive rates is the referral to follow-up of newborns with abnormal results due to heterozygosity (carrier status), a situation not uncommon for disorders such as

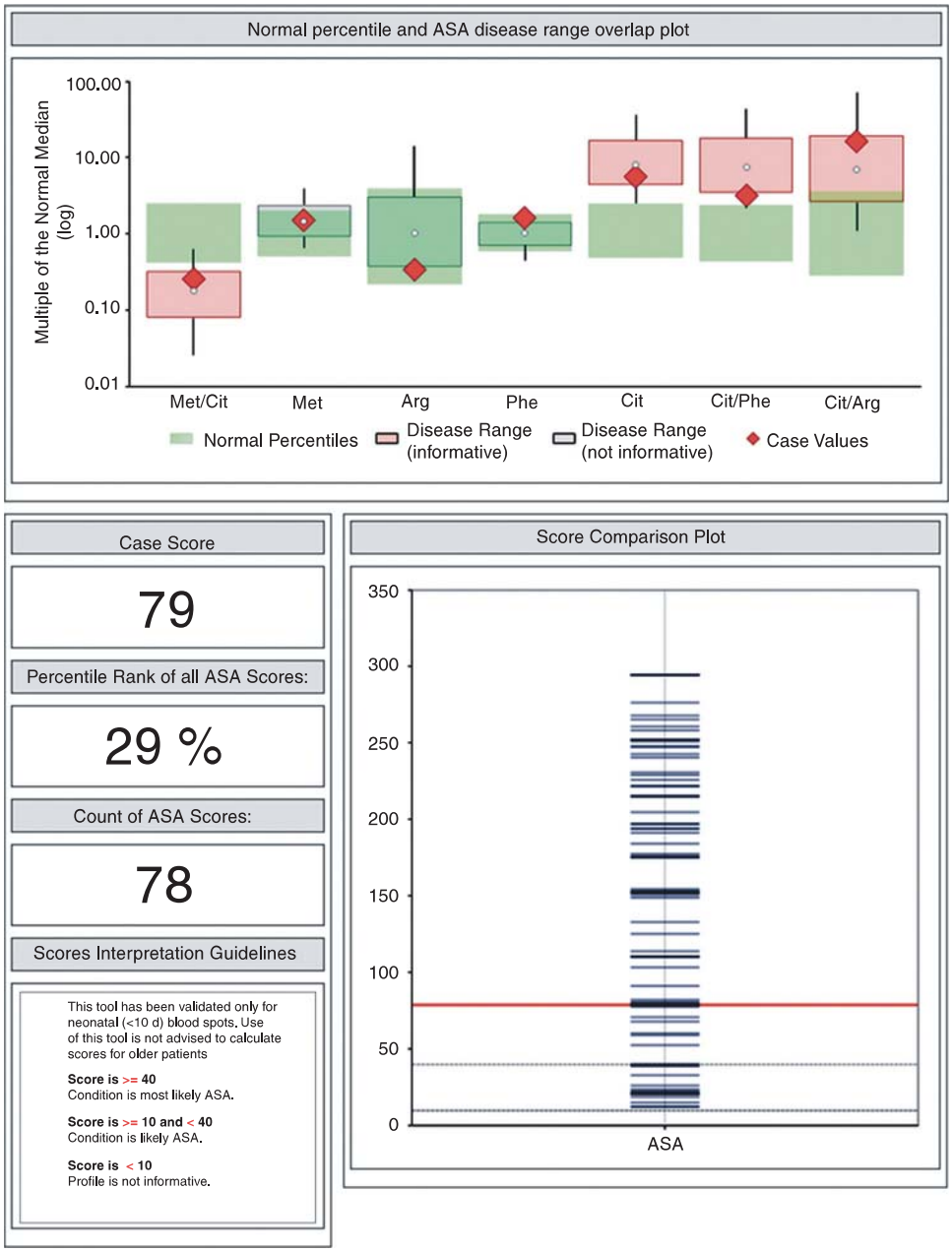

Figure 1 Partial display of the tool for argininosuccinic acid (ASA) lyase deficiency (two of the three panels; see Supplementary Material online for an example of the data-entry panel). This case was considered not informative on the basis of a cutoff for citrulline set inappropriately high. The top panel is an overlay graph of normal population, disease range, and the values entered to calculate a score. All values are expressed as μmol/l and converted to multiples of the normal median on a log scale. The bottom panel shows the calculated score, the percentile rank comparison to all available scores and the case count along with a graphic display of all available scores for the chosen condition, and a summary of interpretation guidelines.

medium-chain acyl-CoA dehydrogenase deficiency[13] and very-long-chain acyl-CoA dehydrogenase deficiency.[14] In both conditions, energy depletion due to prolonged labor and delivery may trigger the transient appearance of a biochemical phenotype mimicking affected status. Interpretive tools can facilitate the identification of carriers and consequently reduce the number of cases requiring follow-up. At the same time, use of these tools could prevent at least some of the false-negative events determined by cutoff values set inappropriately, as mentioned above, but they are not likely to recognize cases with completely uninformative biochemical phenotypes.[10–12,15] Although several acylcarnitine species could be informative for the evaluation of these conditions,[6] the most widely used markers are octanoylcarnitine (C8) and tetradecenoylcarnitine (C14:1), respectively. Figure 2a shows the distribution of paired C8 and C14:1 concentrations in four groups of cases: two with medium-chain acyl-CoA dehydrogenase deficiency (affected and carriers) and two with very-long-chain acyl-CoA dehydrogenase deficiency

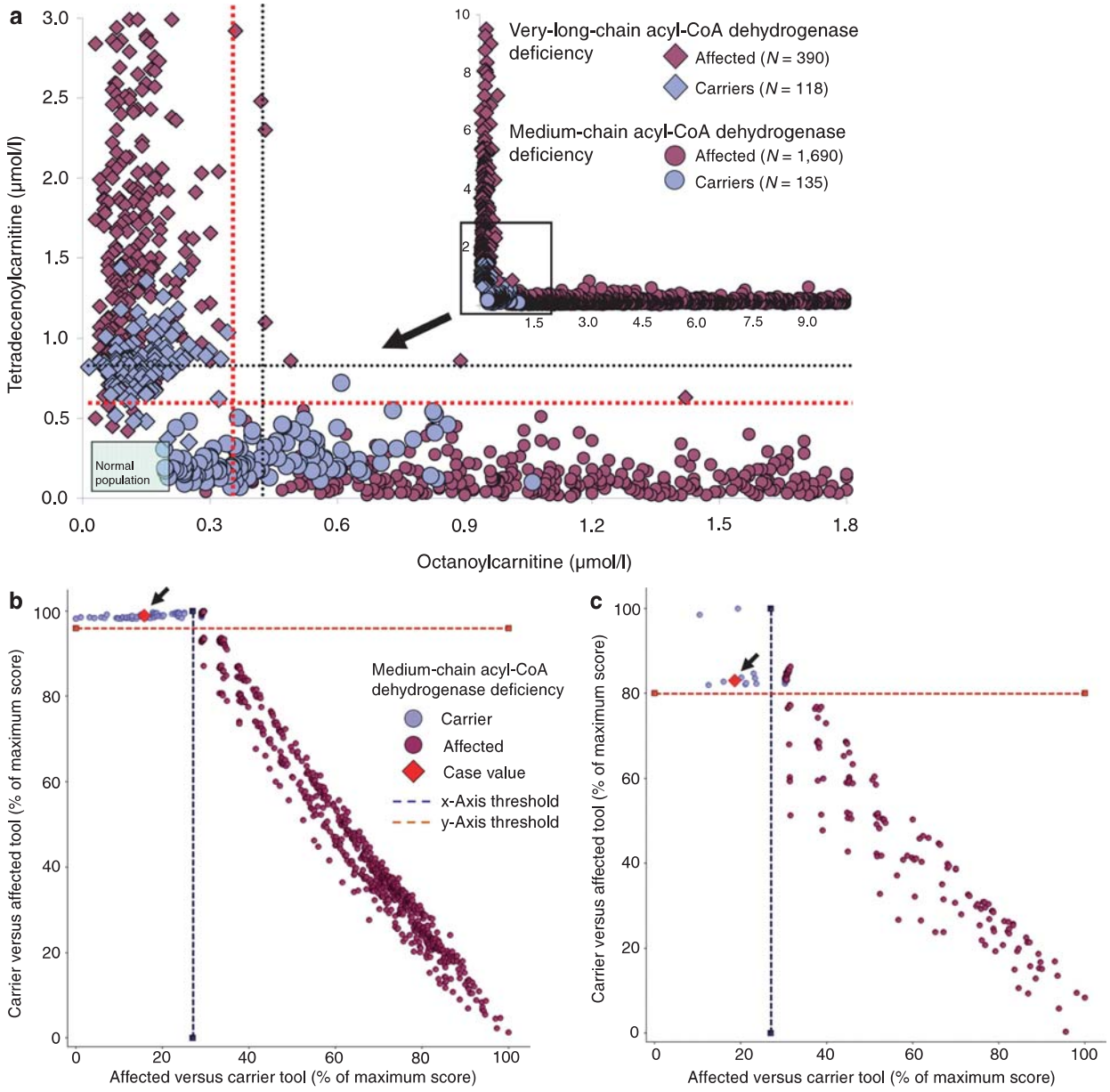

Figure 2 Scatter plots of acylcarnitine results and of condition scores. (**a**) Scatter plot of C8 and C14:1 in four conditions: medium-chain acyl-CoA dehydrogenase (MCAD) deficiency, heterozygote carriers of MCAD deficiency, very-long-chain acyl-CoA dehydrogenase (VLCAD) deficiency, and heterozygote carriers of VLCAD deficiency. The number of cases included in the figure are shown in the insert in the upper right corner, which shows a wider range of values in affected patients (highest values for C8 and C14:1 are 61.8 µmol/l and 13.1 µmol/l, respectively). The horizontal and vertical red dotted lines correspond to the median cutoff value among all laboratories (C8 0.35 µmol/l, *N* = 119; C14:1 0.60 µmol/l, *N* = 113). The horizontal and vertical black dotted lines correspond to the median of the ranges in the two carrier groups (C8 0.44 µmol/l, *N* = 147; C14:1 0.84 µmol/l, *N* = 123). (**b**) Dual scatter plot comparing the scores of MCAD deficiency (dark circles) and MCAD-deficiency heterozygote carriers (light circles). The dotted lines shown as x-axis and y-axis thresholds define the quadrants of the plot where a combined score (x-axis: <27%, y-axis: >96%; x-axis: >27%, y-axis: <96%) is consistent with carrier status and affected status, respectively. The upper right quadrant defines the small area where a combined score is not informative to discriminate carrier vs. affected (hence to be resolved by biochemical and molecular testing); the lower left quadrant is consistent with normal status. (**c**) Dual scatter plot comparing the scores of VLCAD deficiency (dark circles) and VLCAD-deficiency heterozygote carriers (light circles). The dotted lines shown as x-axis and y-axis thresholds define the quadrants of the plot where a combined score (x-axis: <25%, y-axis: >85%; x-axis: >25%, y-axis: <85%) is consistent with carrier status and affected status, respectively. Symbols are the same as in panel **b**. The upper right quadrant defines the area where a combined score is not informative to discriminate carrier vs. affected (to be resolved by biochemical and molecular testing); the lower left quadrant is consistent with normal status.

(affected and carriers). The figure also shows the median values of the two carrier ranges; both values are clearly above the median of all active cutoff values in the collaborative project. These data illustrate how common it may be to encounter an abnormal result due to heterozygosity, a dilemma that cannot be ignored by increasing the cutoff above the carrier range.[16]

The clinical utility of the two conditions tools and, when paired appropriately, of the dual scatter plot is illustrated in panels **b** and **c** of Figure 2. They show the scores of two cases, each generated by a tool based on the same markers but designed to recognize the differences between a target condition (affected) and a secondary condition (carrier). A red diamond symbol marks the location of the combined scores of a medium-chain acyl-CoA dehydrogenase deficiency carrier with a concentration of C8 exactly at the median of the carrier range, Figure 2c does the same for a very long-chain acyl-CoA dehydrogenase deficiency carrier. Cases with values below the median generate scores that are even more segregated, suggesting that, as in the opposite scenario of false-negative events described previously, in at least half of these cases referred to follow-up the cost of unnecessary tests and a variety of unfavorable outcomes[17] could have been prevented. Furthermore, the application of the same postanalytical process to acylcarnitine profiles generated in vitro under controlled circumstances with the fatty-acid probe assay[18] results in a complete separation between the two groups (data not shown). A systematic use of this tool to integrate biochemical and enzymatic results in cases with inconclusive genotyping results has the potential to resolve existing differences of opinion regarding the proper way to follow up an abnormal newborn screening result.[19,20] The in vitro work is beyond the scope of this report and will be published separately (E.H. Smith, D. Matern, et al., unpublished data).

The impact of this objective, evidence-driven approach to the interpretation of laboratory results could be substantial. As an example, Figure 3 shows a longitudinal summary of the performance metrics of newborn screening by tandem mass spectrometry in Minnesota over the period 2002–2010. Minnesota has been the first adopter of all quality-improvement tools made available to the participants of the collaborative project since 2005. The first panel shows the number of true-positive cases per year normalized per 100,000 births. Cases of Hmong ethnicity with 2-methylbutyryl-CoA dehydrogenase deficiency[21] (2–19 new cases per year) were not included in this metric to eliminate the bias of a common disorder in an overrepresented minority. The other two panels show a trend over time of sustained improvement of two performance metrics as described previously,[5] both greatly exceeding the proposed targets of adequate performance (false-positive rate: 0.30%; positive predictive value: 20%).

DISCUSSION

The primary objective of the Region 4 collaborative project is to promote improvement of laboratory quality of newborn

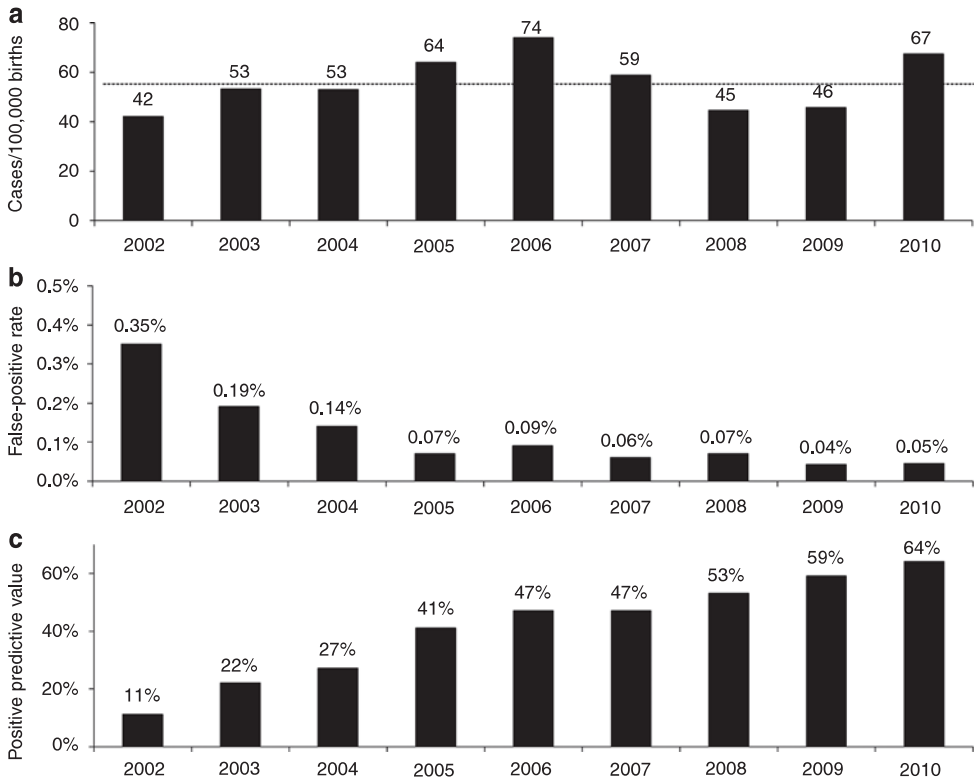

Figure 3 Performance metrics of expanded newborn screening by tandem mass spectrometry in Minnesota, 2002–2010. The birth volume is between 68,000 and 74,000 per year. As result of an ongoing public–private partnership, testing by tandem mass spectrometry was transferred from the Minnesota Department of Health to the Mayo Clinic College of Medicine in June 2004. (**a**) Number of true-positive cases per year normalized to 100,000 births. The dotted line indicates the average (55.8/year/100,000 births). (**b**) Trend of false-positive rate. This metric is expressed as the proportion of positive tests in subjects proven by follow-up evaluation not to have one of the conditions targeted by the Minnesota program.[7] (**c**) Trend of positive predictive value. This metric is expressed as the probability that a newborn is affected with a condition when restricted to cases with a positive test.[7] Between June 2004 and December 2010, no false-negative events were brought to the attention of the program with respect to a condition included in the uniform panel.[3]

screening by tandem mass spectrometry. Dealing with rare conditions of undetermined prevalence, a database of meaningful clinical utility could be produced only through an unprecedented level of cooperation and collaboration on a global scale. The database has led to a new and original type of interpretive tool to achieve reduction in both false-negative events and false-positive outcomes. More traditional statistical methods for separation of cases from noncases, such as likelihood ratio methods and discriminant analysis,[22] are not appropriate in the current situation because they assume a multivariate normal distribution of the analyte values in the cases that is not observed. Although many of the detected disorders arise from mutations in a single gene, the variability of the mutations and the extent of the corresponding phenotypic variation are unknown. As such, most of the disease populations are complex mixtures that cannot be modeled with simple parametric distributions. Reliable information for some of the required characteristics, for example, the prevalence of the disease and the complexity of the differential diagnosis needed for a majority of the informative markers, is also lacking. A further disallowing complexity is the number of covariance parameters to be estimated, which vastly exceeds the number of cases of all but the most common of the disorders, making the parametric distribution subject to significant bias.

The lack of traditional analyte cutoff values may seem counterintuitive for reporting quantitative laboratory test results on which binary decisions will be based. However, the basic tenet of this multivariate pattern-recognition software is that an abnormal result is not defined exclusively by a deviation from a statistical definition of normal. The software also evaluates how consistent a result is with the analyte disease range established separately for each condition, an assessment that is novel and more informative than a traditional "one size fits all" cutoff value, and is made possible by a database of true-positive cases of unprecedented size. Another distinctive advantage of the postanalytical tools is the opportunity to calibrate any decision with an element that has not been taken full advantage of so far, which is the degree of overlap between normal population and disease range.

The interpretive tools first became available in January 2009. A conservative estimate of the utilization of the versions based on static spreadsheets is on the order of tens of thousands of downloads; more than 17,000 page views have been recorded since the initial release of the online tools (23 March 2011). The feedback from a diverse spectrum of users, laboratorians, and clinicians has been consistently positive, with indications that these tools are now used in clinical practice on a regular basis and indeed are effective, providing independent verification of the single-site evidence shown in **Figure 3**. A sustained trend of constant improvement is significant because cost–benefit analysis, expense management, and optimization of resource utilization are high priorities in these times of increasing financial constraints, and the public health infrastructure is not exempt from the demand for reducing the cost of health-care services. Future recommendations to expand the uniform newborn screening panel with the addition of more conditions[23–25] will raise this pressure even more.

This approach is flexible by design and certainly not limited to amino acids and acylcarnitines. It has already been successfully applied to other multianalyte profiles currently used as either primary or second-tier newborn screening tests, for example, for the interpretation of steroid profiles in congenital adrenal hyperplasia[26] and of C_{20}-C_{26} lysophosphatidylcholine species in X-linked adrenoleukodystrophy and other peroxisomal disorders.[27] The availability of more diverse applications is limited only by the gathering of sufficient data of the normal population and of patients affected with the target condition(s).

The software continues to incorporate improvements suggested by users, for example, the ability to customize the pool of percentiles and affected cases relied on to calculate scores. Users have the option to display scores based on subgroups of cases, either their own cases, those belonging to a specific country or, in the future, contributed by laboratories having the closest participant profile in terms of analyte percentiles in the normal population. Additional functions scheduled to be released in the near future are an "all conditions" tool (an unrestricted evaluation of full amino acid and acylcarnitine profiles to suggest any possible diagnosis) and interfaces to download entire batches of raw data from existing commercial software. Additional applications unrelated to newborn screening will become routinely available to span a broad spectrum of either clinical or research endeavors. This evidence-based approach could add substantial value to patient care by providing a comprehensive interpretation of complex laboratory profiles driven by cumulative/multisite evidence and by objective peer comparison.

SUPPLEMENTARY MATERIAL

Supplementary material is linked to the online version of the paper at http://www.nature.com/gim

ACKNOWLEDGMENTS

Additional public health and private programs have contributed data to the collaborative project, but no individual affiliated with these sites met the authorship criteria for this report. US contributors are from Alabama, Arizona, Arkansas, Colorado, Delaware, Hawaii, Illinois, Maine, Mississippi, Montana, Nebraska, New Hampshire, New Jersey, North Dakota, Ohio, Oklahoma, Pennsylvania, Puerto Rico, Rhode Island, South Dakota, Texas, Utah, Vermont, Virginia, West Virginia, and Wyoming. Partial contributions from international programs were from Argentina, Australia, Austria, Belgium, Brazil, Bulgaria, Canada, Chile, China, Croatia, Czech Republic, Denmark, Germany, Greece, Hungary, India, Ireland, Italy, Japan, Malaysia, Mexico, Norway, Poland, Qatar, Saudi Arabia, South Africa, Spain, Sweden, Switzerland, Turkey, United Arab Emirates, and the United Kingdom.

This work was supported by a grant (U22MC03963) to the Region 4 Genetics Collaborative from the Health Resources and Service Administration of the Maternal and Child Health Bureau Cooperative Agreement; by contracts from the Eunice Kennedy Shriver National Institute of Child Health and Human

Development, National Institutes of Health, Department of Health and Human Services (contract HHSN275201000017C) and the Newborn Screening Translational Research Network (subcontract HHSN275200800001C 01); and by the T. Denny Sanford Professorship fund, Mayo Clinic College of Medicine.

REFERENCES

1. Wilcken B, Wiley V, Hammond J, Carpenter K. Screening newborns for inborn errors of metabolism by tandem mass spectrometry. *N Engl J Med* 2003;348:2304–2312.
2. Watson MS, Mann MY, Lloyd-Puryear MA, Rinaldo P, Howell RR. Newborn screening: Toward a uniform screening panel and system—Executive summary. *Genet Med* 2006;8(suppl):1S–11S.
3. Koppaka R. Ten great public health achievements—Unites States 2001–2010. *MMWR* 2011;60:619–623.
4. Sebelius K. Response by the HHS Secretary to the February 25, 2010 and November 22, 2009 letters. http://www.hrsa.gov/heritabledisorderscommittee/correspondence/ response5_21_2010.pdf.
5. Rinaldo P, Zafari S, Tortorelli S, Matern D. Making the case for objective performance metrics in newborn screening by tandem mass spectrometry. *Ment Retard Dev Disabil Res Rev* 2006;12:255–261.
6. McHugh DM, Cameron CA, Abdenur JE, et al. Clinical validation of cutoff target ranges in newborn screening of metabolic disorders by tandem mass spectrometry: a worldwide collaborative project. *Genet Med* 2011;13:230–254.
7. Howell RR. Quality improvement of newborn screening in real time. *Genet Med* 2011;13:205.
8. Downs SM, van Dyck PC, Rinaldo P, et al. Improving newborn screening laboratory test ordering and result reporting using health information exchange. *J Am Med Inform Assoc* 2010;17:13–18.
9. Sreenath Nagamani SC, Erez A, Lee B. Argininosuccinate lyase deficiency. In GeneReviews (Internet). Pagon RA, Bird TD, Dolan CR, et al. (eds). University of Washington: Seattle, 1993. http://www.ncbi.nlm.nih.gov/sites/GeneTests/. Accessed 3 June 2011.
10. Gallagher RC, Cowan TM, Goodman SI, Enns GM. Glutaryl-CoA dehydrogenase deficiency and newborn screening: retrospective analysis of a low excretor provides further evidence that some cases may be missed. *Mol Genet Metab* 2005;86:417–420.
11. Puckett RL, Lorey F, Rinaldo P, et al. Maple syrup urine disease: further evidence that newborn screening may fail to identify variant forms. *Mol Genet Metab* 2010;100:136–142.
12. Sarafoglou K, Matern D, Redlinger-Grosse K, et al. Siblings with mitochondrial acetoacetyl-CoA thiolase deficiency not identified by newborn screening. *Pediatrics* 2011;128:e246–e250.
13. Matern D, Rinaldo P. Medium-chain acyl-CoA dehydrogenase deficiency. In GeneReviews (Internet). Pagon RA, Bird TD, Dolan CR, et al. (eds). University of Washington: Seattle, 1993. http://www.ncbi.nlm.nih.gov/sites/GeneTests/. Accessed 3 June 2011.
14. Leslie ND, Tincle BT, Strauss AW, Shooner K, Zhang K. Very long chain acyl-Coenzyme A dehydrogenase deficiency. In GeneReviews (Internet). Pagon RA, Bird TD, Dolan CR, et al. (eds). University of Washington: Seattle, 1993. http://www.ncbi.nlm.nih.gov/sites/GeneTests/. Accessed 3 June 2011.
15. Sahai I, Bailey JC, Eaton RB, Zytkovicz T, Harris DJ. A near-miss: very long chain acyl-CoA dehydrogenase deficiency with normal primary markers in the initial well-timed newborn screening specimen. *J Pediatr* 2011;158:172; author reply 172–172; author reply 173.
16. Maier EM, Pongratz J, Muntau AC, et al. Validation of MCADD newborn screening. *Clin Genet* 2009;76:179–187.
17. Waisbren SE, Albers S, Amato S, et al. Effect of expanded newborn screening for biochemical genetic disorders on child outcomes and parental stress. *JAMA* 2003;290:2564–2572.
18. Matern D. Acylcarnitines, including in vitro loading tests. In: Blau N, Duran M, Gibson KM (eds). Laboratory Guide to the Methods in Biochemical Genetics. Springer: Berlin, Germany, 2008:171–206.
19. Arnold GL, Van Hove J, Freedenberg D, et al. A Delphi clinical practice protocol for the management of very long chain acyl-CoA dehydrogenase deficiency. *Mol Genet Metab* 2009;96:85–90.
20. Kronn D, Mofidi S, Braverman N, Harris K; Diagnostics Guidelines Work Group. Diagnostic guidelines for newborns who screen positive in newborn screening. *Genet Med* 2010;12(12 suppl):S251–S255.
21. Matern D, He M, Berry SA, et al. Prospective diagnosis of 2-methylbutyryl-CoA dehydrogenase deficiency in the Hmong population by newborn screening using tandem mass spectrometry. *Pediatrics* 2003;112(1 Pt 1):74–78.
22. Johnson RA, Wichern DW. Applied Multivariate Statistical Analysis, 3rd edn. Prentice Hall: Englewood Cliffs, NJ, 1992:552–556.
23. Green NS, Rinaldo P, Brower A, et al.; Advisory Committee on Heritable Disorders and Genetic Diseases in Newborns and Children. Committee Report: advancing the current recommended panel of conditions for newborn screening. *Genet Med* 2007;9:792–796.
24. Perrin JM, Knapp AA, Browning MF, et al. An evidence development process for newborn screening. *Genet Med* 2010;12:131–134.
25. Calonge N, Green NS, Rinaldo P, et al.; Advisory Committee on Heritable Disorders in Newborns and Children. Committee report: Method for evaluating conditions nominated for population-based screening of newborns and children. *Genet Med* 2010;12:153–159.
26. Lacey JM, Minutti CZ, Magera MJ, et al. Improved specificity of newborn screening for congenital adrenal hyperplasia by second-tier steroid profiling using tandem mass spectrometry. *Clin Chem* 2004;50:621–625.
27. Hubbard WC, Moser AB, Liu AC, et al. Newborn screening for X-linked adrenoleukodystrophy (X-ALD): validation of a combined liquid chromatography-tandem mass spectrometric (LC-MS/MS) method. *Mol Genet Metab* 2009;97:212–220.

Manuscript # GIM-D-11-00263R3
Authors: Marquardt G, Currier R, McHugh DMS, et al.
Title: Enhanced interpretation of newborn screening results without analyte cutoff values

Online Supplemental Material

MATERIALS AND METHODS

Tool builder

Tools are created according to a stepwise process, called the tool builder, which is accessible on the website to a single user per participating site. This restriction stems from the concern that multiple individuals may modify or delete tools prepared by another user without adequate internal communication. Once released, a site-specific tool can be used instead of the general tool but only by users affiliated with the same site, the default tool for any given condition remains accessible to all participants. Separate tools for the same condition are available to account for differences between derivatized and underivatized markers, when applicable.

Functions are selected according to the choices shown in Table 1 (Suppl). A description of the more complex functions is provided below, starting with the scoring type, which could be decreasing, plateau, or increasing. All three scoring strategies are applied in two phases. In the first phase, which is the same in all three modalities, points are given starting at 1 and going to 6 in increments of one. The assignment of points begins with the lowest target percentile (from a total of 12 selected for scoring purposes, see Table 2 Suppl.) in the disease range that is beyond the 99th percentile of the normal range. In the decreasing strategy, after a score of 6, the score for the next target percentile exceeded by the result is decreased by one and so forth until a score of 1 is reached again. The total score for the analyte is then the sum of the points assigned for each

percentile exceeded by the result. In a best case scenario, one where there is no overlap between normal population and disease range, when using the decreasing mode the maximum cumulative score for an analyte is 36. In the least informative scenario, the lowest possible score for a marker with a value >80th percentile of the disease range is 10 (in a situation where the 40th percentile of the disease range is below the 99th percentile of the normal population). In other words, the degree of overlap rules that no points are added to the total for the analyte up to the 40th percentile. The plateau strategy differs from the first one in that after accruing a score of 6 all subsequent percentiles exceeded by the result are assigned the same score of 6 (Table 2-Suppl). The highest possible score becomes 57, the lowest is still 10 for the case where the 40th percentile of the disease range is below the 99th percentile of the normal population. Finally, the increasing strategy continues beyond a score of 6 to a maximum of 12, the highest possible cumulative score for one analyte is 78. Regardless of which strategy is used, results falling into the highest quintile of disease ranges above the normal population (>80th percentile) and into the lowest one of disease ranges below the normal population (<20th percentile) do not contribute to the score. Alternative scoring strategies are desirable because they could be used to underscore differences between clusters of patients (for example, reflecting segregation according to established genotype to phenotype correlations, responsiveness to treatment, and differences in age of onset). [1-2] The software to create a tool is complemented by a parallel function, described as the tool tuner, which allows to assess impact and differences of alternative criteria between an existing tool (released) and a new version still under development (not released). The tuner function is also valuable to investigate retrospectively the basis for either false positive and false negative results, suggesting potential corrective action using outlier rules in the tool builder process.

The establishment of disease ranges has led to the recognition that the pool of informative markers for a given condition may differ vastly in clinical significance. Therefore, the tools correct raw scores by a calculated factor that reflects the comparative significance of all analytes under consideration. For each analyte the extent of the disease range not overlapping with the normal range is calculated (excluding the last quintile at the opposite end of the normal population), and combined with all others. The percentage of the total is incorporated proportionally in the correction factor of each analyte. For example, an analyte that amounts to 40 percent of the calculated total is assigned a correction factor of 1.40. In other words, the raw score is increased by 40 percent.

The inclusion in a tool of individual high and low informative markers is at the discretion of the user and is accomplished by selecting check boxes sorted according to the degree of overlap between normal population and disease range. When applicable, the software enables the choice between derivatized and underivatized markers to account for analytical differences between laboratories. When more than one condition is included, the discriminating power of a tool is derived from the analysis of any degree of separation between disease ranges. Figure 1S (Suppl) shows a side by side comparison of the disease ranges of a generic analyte in two model conditions, labeled as A and B. The first percentile of the disease range of condition A matches the median of condition B, the 99^{th} percentile of the disease range of condition B is equal to the median of condition A. Even though half of each disease range overlaps with the other one, the software actually takes advantage of this behavior to achieve an informative differential diagnosis as shown in the three examples (1, 2, and 3). In example 1, the value is treated differently depending on which of the two conditions is the primary target. In a tool for condition A, the observed result translates in an added score (rule 1, see below how influence is

determined). In the opposite scenario, a tool for condition B, it will actually trigger a reduction of the score (rule 3). In example 2 the score is accrued in both tools only on the basis of the degree of separation from the normal population, but it will be greater in condition B as the value corresponds to a much higher percentile of the disease range (condition A: ~10^{th} percentile; condition B: ~90^{th} percentile). Finally, the third example will generate a reduction of the score in the tool for condition A (rule 2), and an increase in the tool for condition B (rule 4). The impact of these rules is based on a numerical value equal to the maximum attainable analyte score of the chosen scoring strategy (decreasing, plateau, increasing) multiplied by the number of informative markers incorporated in the tool, and is further modified by a correction factor. A choice is possible between the weighted calculation described previously and two others (described as percentage and percentile) that are based on the redefinition of the disease range as the interval of values entirely outside of the normal population. The screening result is then expressed as either a percentage or as a percentile rank, and the correction factor is adjusted accordingly. For example, a value corresponding to the 28% of the modified disease range (or to the 28th percentile) receives a correction factor of 1.28 (1.0 + 0.28). The highest possible correction factor is 2.0 (1.0 + 1.0). When this process is applied in parallel to all informative markers it determines a cumulative effect which overcomes significantly the overlap at the level of individual analytes when they are compared between related conditions.

In the evaluation of a single case, the score of individual tools is not influenced by the inclusion of markers that are informative for multiple conditions. However, the inclusion of the same ratio in multiple tools becomes problematic when scores are calculated simultaneously, an option called “all conditions tool”. For this reason, the differentiator and outlier functions are in place to establish rules to prevent conflicting, and potentially confusing, scores. For example, the

tool for remethylation disorders [3] is prevented from assigning points in cases where the low methionine/phenylalanine ratio is informative because of an elevated concentration of phenylalanine, not a reduced concentration of methionine.

As the next step, filters are applied to exclude cases with an analyte result either below or above a selectable value or a percentile of the disease range. In view of the wide utilization of the tools in routine newborn screening settings, this function is needed because new submitted cases are immediately incorporated in the disease ranges. In consideration of the worldwide location of users and a virtually around the clock submission cycle, the consistency and reliability of the tools must be protected by a mechanism to prevent disruptions triggered by data entry errors and incorrect diagnoses. In these situations changes to the disease ranges due to automatic inclusion of all submitted cases could be significant before anomalies or errors are detected and corrected, especially when conditions still have a relatively small count (<50). Filters can be set for a value or a disease range percentile (for example first percentile and 99th percentile), with the option to include or not the filtered values in the disease range and also to apply the same filter to all secondary conditions included in the tool. Once a single analyte value has been filtered, no score is calculated for that case. The choice of filters is supported by access to a tabular summary of the disease ranges of all analytes, and by the option to display the top 50 high and low values of any analyte for each condition. The outliers to be considered for exclusion are shown as actual value, standard z-score [(value-mean of the disease range)/standard deviation],[4] percentile of the disease range, multiple of the median of the normal population, and multiple of the median of the disease range. If desired, the entire set of data of any case is available on a separate tab for a more in depth evaluation. This composite profile empowers extensive flexibility to select a threshold, if one is deemed to be either clinically necessary or desirable as a security feature to

ensure consistency and reliability of the tools.

The impact and run functions are the testing and validation environment of the tool builder and allow a stepwise assessment of differentiators, correction factors, and filters as previously selected either individually or in any combination. When a score is calculated on the run page it is possible to display the contribution of each analyte of any case of the target condition and of the secondary conditions, including the equation to determine the correction factors. To achieve consistency among tools comparing two or more conditions, the range of scores is adjusted by applying a minimum-maximum normalization.[5] This calculation transforms each value so that the maximum result for the column is 100 and the minimum is zero. Each result is calculated by subtracting from the score the lowest of all scores, dividing it by the range of values (highest minus lowest), and multiplying by 100. This preserves the relative distance between values. Alternatively, the formula [(z-score x 100) +500] can be applied. This transformation expands the range of scores so that 95% of scores fall between 300 and 700, with the addition of 500 functioning to shift scores reduced by differentiators to a positive number. The rationale for this final normalization by either method is to keep all tools on a comparable scale, a feature particularly important for the "all conditions" tool. This functionality is under validation as a potential primary mechanism for evaluation of whole batches of cases uploaded electronically to the website.

As the final step, condition-specific guidelines are provided for score interpretation. Each tool clearly indicates the score below which the profile is deemed to be not informative (between zero and the lowest score of a known case), and incremental thresholds indicating when a score is either "possibly", "likely", or "most likely" to be consistent with a biochemical diagnosis of the target condition. When applicable, the guidelines suggest performing a second tier test[6-7] or

reflexing to an available dual scatter plot. The highest score considered to be uninformative becomes the only cutoff needed for the target condition, replacing all analyte cutoff values and their inherent limitations and selection biases. Dual scatter plots are released with X- and Y-axis thresholds drawn to separate the two conditions under consideration, interpretation guidelines are based on the location of a combined score in one of the four quadrants defined by such lines.

When a tool is accessed manually, users make a selection from a menu of available choices and are presented with a window (see Figure 2A-Suppl) in which the required values are entered to calculate the score. The data entry window shown in the figure is for the condition carnitine palmitoyltransferase 1A deficiency, a fatty acid β-oxidation disorder.[8] Analytes are displayed in three groups: low markers, differentiators, and high markers. The actual tool includes three panels: the first panel (Figure 2B-Suppl) is a summary of the relevant percentiles of the normal population (first percentile for low markers and 99th percentile for differentiators and high markers), the degree of overlap in percent of the target disease population (i.e., the percent of the disease range to which either the first percentile or 99th percentile of the normal population corresponds), the three percentiles of the disease range closest to the normal population (99th percentile, 95th percentile, and 90th percentile for low markers; first percentile, fifth percentile, and 10th percentile for differentiators and high markers) and the median of the disease range. A darker shade over the percentile data visualizes the extent of overlap between normal population and disease range. After the score has been calculated, another column will appear (not shown) with all the analyte values imported from the data entry window and any pertinent ratios calculated from the case analyte values. It is notable that a number of routinely calculated ratios may actually reach a threshold of clinical significance in a wide spectrum of conditions but frequently are not appreciated as informative elements of the expected

biochemical phenotype. For example, the propionylcarnitine/palmitoylcarnitine ratio has not been mentioned before our report as an informative marker for carnitine palmitoyltransferase 1A deficiency, [8] despite the evidence that approximately 75 percent of affected cases have an elevated ratio.

The third panel (see Figure 1 of published article) of tools, excluding the dual scatter plot format, summarizes the calculated score, the percentile rank, the number of cases available to calculate scores (partial sets of results are still included in disease ranges), the interpretation guidelines according to a standardized format (not informative, possibly, likely, most likely), and a graph of all scores superimposed to dotted lines matching the guidelines and the calculated score. Users can click an icon to save the report as a portable document format file, [9] a practice which is desirable for documentation purposes because the tools are by design constantly evolving. The score for the same set of results may change in a matter of days if more cases have been added since the previous use of the tool. To facilitate documentation, the printable file automatically includes a header with the name of the primary condition, the unique version number of the tool and the date it was created, the type of tool (single, dual, multiple), and access (available to all or only to users of a single site). The header also displays the date and time the tool was printed, and the name and site affiliation of the user who entered the data to calculate a score.

REFERENCES

1. Smith EH, Thomas C, McHugh D, et al. Allelic diversity in MCAD deficiency: the biochemical classification of 54 variants identified during 5 years of ACADM sequencing. Mol Genet Metab. 2010;100(3):241-250.

2. Ensenauer R, Fingerhut R, Maier EM, et al. Newborn screening for isovaleric acidemia using tandem mass spectrometry: data from 1.6 million newborns. Clin Chem. 2011;57(4):623-626.

3. Tortorelli S, Turgeon CT, McHugh DMS, et al. Two-tier approach to the newborn screening of methyleneteratrahydrofolate reductase deficiency and other re-methylation disorders by tandem mass spectrometry. J Pediatr 2010;157:271-275.

4. Jackson SL. Research methods and statistics: A Critical thinking approach. 3rd edition. Belmont, CA.: Wadsworth, 2009.

5. Normalize wizard. www.predixionsoftware.com. Accessed June 3, 2011

6. Matern D, Tortorelli S, Oglesbee D, Gavrilov D, Rinaldo P. Reduction of the false positive rate in newborn screening by implementation of MS/MS-based second tier tests: The Mayo Clinic experience (2004-2007). J Inherit Metab Dis. 2007;30(4):585-592.

7. Turgeon CT, Magera MJ, Cuthbert CD, et al. Determination of total homocysteine, methylmalonic acid, and 2-methylcitric acid in dried blood spots by tandem mass spectrometry. Clin Chem. 2010;56(11):1686-1695.

8. Bennett MJ, Narayan SB, Santani AB. Carnitine palmitoylcarnitine 1A deficiency. In GeneReviews [Internet]. Pagon RA, Bird TD, Dolan CR, et al, editors. Seattle (WA): University of Washington, Seattle; 1993-.http://www.ncbi.nlm.nih.gov/sites/GeneTests/ Accessed June 3, 2011.

9. Adobe Corporation. http://www.adobe.com/ Accessed June 3, 2011.

Table 1 (Suppl) – Functions and choices available in the tool builder

Function	*Choices*	*Default setting*	*Rationale*
Access	Available to all sites Available to a single site only	Available to all sites	Single site tools enable customization to match pre-analytical characteristics (derivatized vs. underivatized method) and analyte selection
Type	One condition Two conditions Multiple conditions	One condition	Informative/not informative score for one condition or differential diagnosis between two or more conditions
Scoring strategy	Increasing Plateau Decreasing	Increasing	Alternative options modulate the numerical distribution of scores at the high end. A decreasing strategy (see text) is indicated for conditions with many (>10) informative markers, an increasing strategy is indicated for conditions with <5 informative markers
Correction factors	Weighted Percentile Percentage	Weighted	Weighted correction factors are condition-specific, the others are customized for each case. All are calculated from the degree of overlap between the normal population and the disease range of each informative marker in a given condition
Score type	Regular Z-score Min-Max normalization	Regular	Z-score and Min-Max normalization are optional methods of score manipulation available to improve the visual display of tools targeting multiple conditions
Informative markers	All analytes and ratios that exceed the threshold of clinical significance	All analytes and ratios with a disease range median outside of the normal population range (high or low)	An abnormal result is not defined exclusively by a deviation from a statistical definition of normal, as in a cutoff-based approach. The selection is broader than using conventional standards, especially for ratios where the primary marker is selected as denominator. Informative markers are also selectable to set differentiator rules and filters (see below)
Additional markers	All analytes and ratios that do not exceed the threshold of clinical significance	All analytes required to calculate an informative ratio	Additional markers are selectable to set outlier rules and filters (see below)
Differentiators (one condition)	Set case score to zero when a primary marker is normal	None selected	While all informative markers contribute to the generated case score, some informative markers are more important than others. These markers are considered "primary" markers. If the value for a primary marker is normal (i.e. within the range of the normal population), then the condition being tested can be ruled out. This is true even if the other informative markers generate a case score that is informative. This function allows the user to designate which, if any, markers are primary and specify the normal population threshold that will cause a case score of zero.

Table 1 (Suppl) – continued

Function	*Choices*	*Default setting*	*Rationale*
Differentiators (two or multiple conditions)	Percentile, value (if < or > add/remove points) Rules	Rule 5	Rule 1: add points above the highest value of all secondary conditions; Rule 2: subtract points below the lowest value of the primary condition; Rule 3: subtract points above the highest value of primary condition; Rule 4: add points below the lowest value of all secondary conditions Rule 5: all of the above
Ratio Outliers (one condition)	Enabled or disabled	Enabled	Many ratios consist of some combination of informative and non-informative markers (informative/non-Informative or non-Informative/informative). The ratio outliers rule prevents awarding points to a ratio if the non-informative value is outside the disease range of the condition. For example, for an informative high ratio where the numerator is non-informative, if the numerator is greater than the 100%ile of the disease range, then the ratio's score is set to zero. Alternatively, if the denominator is the non-informative marker and it is less the 0%ile of the disease range, then the ratio's score is set to zero. The purpose of this function is to prevent awarding points to ratios where the non-informative marker drives the penetration into the disease range, not the informative marker.
Outliers (one condition)	Create rules for non-informative markers	None selected	Points are awarded to a case based on the degree of penetration of informative markers into the condition's disease ranges. However, markers may be informative for several conditions. This can cause scenarios where a marker value can result in several conditions generating an informative score. In these cases it is necessary to identity outlier values amongst the non-informative markers. With an outlier rule, the user can specify that the case score should be set to zero if the non-informative marker value is above or below a specified percentile of the condition's disease range.
Filters	Filter type (%ile, value) < or >	Filter values <1%ile and >99%ile of disease range	To prevent undetected data entry errors or absurd values from altering the disease range of informative markers and consequently impacting the real time performance of a tool
Impact	Select any combination of the rules established above	None selected	To detect any significant outlier that generate a non informative score, and to begin the process of guideline selection (threshold of informative scores)

Table 1 (Suppl) – continued

Function	***Choices***	***Default setting***	***Rationale***
Run	Enter a set of values (or a case ID number) to test the tool with visualization of calculations for each informative marker	None entered	This function also allows to switch and compare scores using other scoring strategies, correction factors, and score types
General Guidelines	A place to include information about the tool, disclaimers, and limitation (see default)	This tool has been validated only for neonatal (<10 days) blood spots. Use of this tool is not advised to calculate scores for older patients.	Free text could be added as header of the interpretation guidelines as deemed clinically indicated
Range Guidelines	(<score) Profile is not informative (<score>) Condition is possibly (< score>) Condition is likely (score>) Condition is very likely	None selected	Guidelines formalize the transition from individual and generic analyte cutoff values (one selected for all conditions potentially related to a marker) to a cumulative, condition-specific threshold of clinical significance

Table 2 (Suppl) - Point assignment by alternative scoring strategies

Percentile exceeded		Mode/Points assigned		
High marker	**Low marker**	**Decreasing**	**Plateau**	**Increasing**
1	99	1		
5	95	2		
10	90	3		
15	85	4		
20	80	5		
25	75	6		
30	70	5	6	7
40	60	4	6	8
50	50	3	6	9
60	40	2	6	10
70	30	1	6	11
80	20	0	6	12
Highest possible score		**36**	**57**	**78**

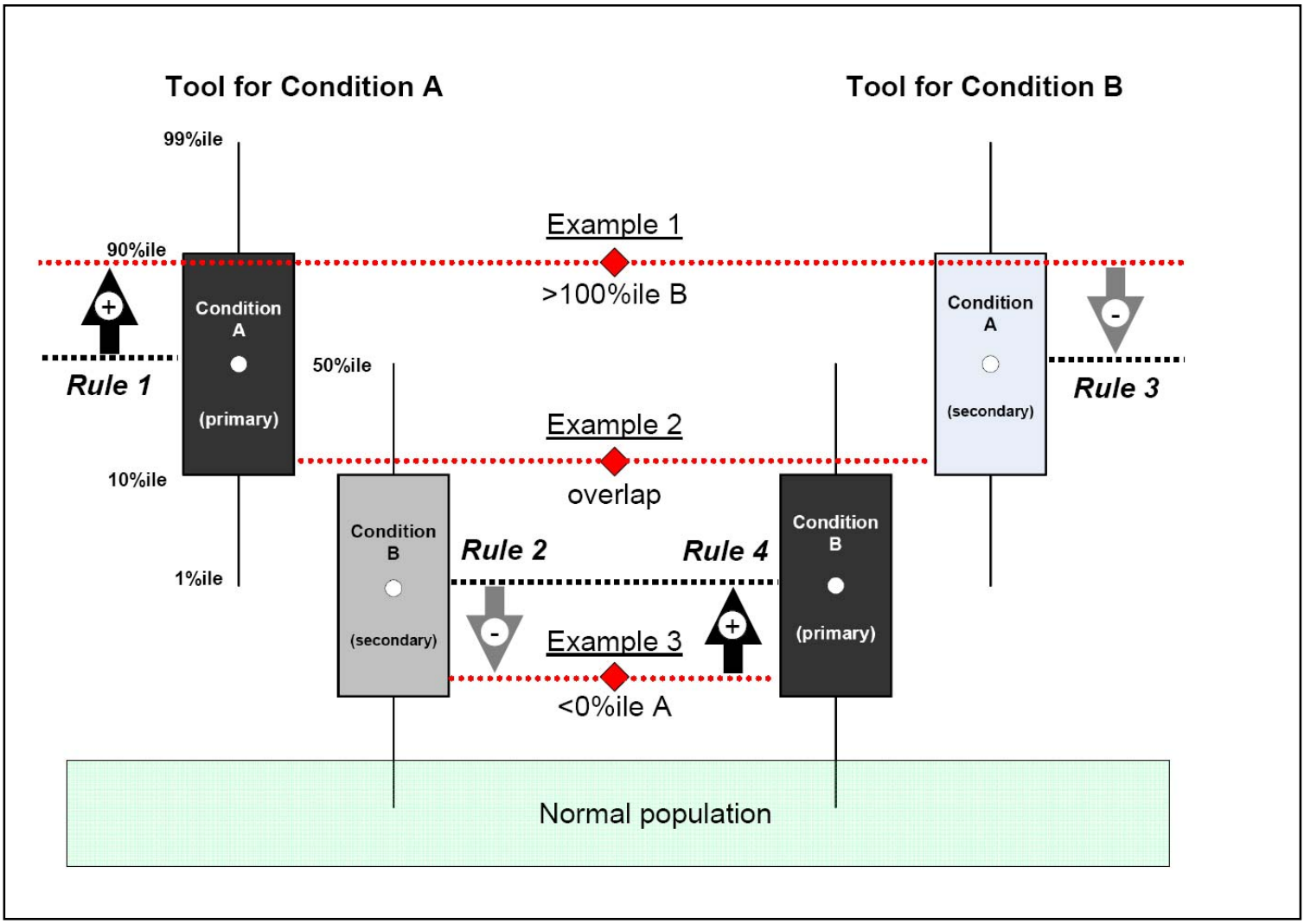

Figure 1 (Suppl) - Examples of the four rules applied to add or subtract points to the raw score of an analyte for a condition.

The same result may lead to significantly different scores for one condition vs. another, and particularly whether it is chosen as the primary condition or as a secondary one. See text for the description of example 1, 2, and 3 scenarios.

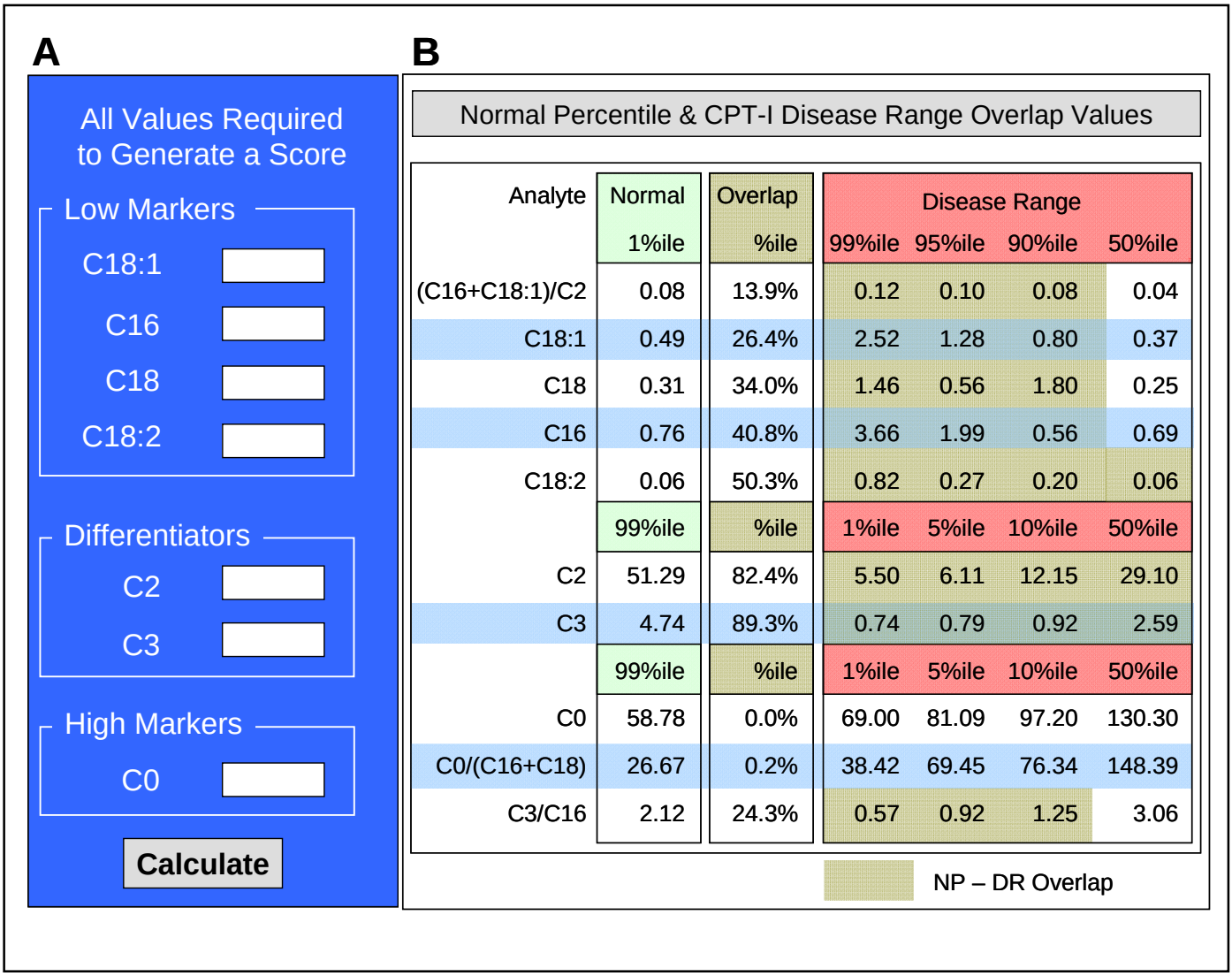

Figure 2(Suppl) - One condition post-analytical tool for carnitine palmitoyltransferase 1A deficiency.

This tool is not inclusive of cases carrying the Northwest native mutation. **A**, data entry window. The unit of results (not shown) is µmol/liter for all analytes. **B**, summary window of a selected percentiles for each informative analyte in the normal population and disease range. Also shown is the degree of overlap between the two ranges. 0% would indicate no overlap (i.e., a complete separation between the two ranges).