Clinical Biochemistry 116 (2023) 7–10

Contents lists available at ScienceDirect

Clinical Biochemistry

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

Short Communication

Machine learning to improve false-positive results in the Dutch newborn screening for congenital hypothyroidism

Kevin Stroek [a], Allerdien Visser [b], Catharina P.B. van der Ploeg [c], Nitash Zwaveling-Soonawala [d], Annemieke C. Heijboer [a,e], Annet M. Bosch [f], A.S. Paul van Trotsenburg [d], Anita Boelen [a], Mark Hoogendoorn [g], Robert de Jonge [h,*]

[a] Endocrine Laboratory, Department of Clinical Chemistry, Amsterdam Gastroenterology, Endocrinology & Metabolism, Amsterdam UMC, University of Amsterdam, Amsterdam, The Netherlands
[b] Department of Clinical Chemistry, Amsterdam Gastroenterology, Endocrinology & Metabolism, Amsterdam UMC, Vrije Universiteit, Amsterdam, The Netherlands
[c] Netherlands Organization for Applied Scientific Research TNO, Department of Child Health, Leiden, The Netherlands
[d] Department of Paediatric Endocrinology, Emma Children's Hospital, Amsterdam Gastroenterology, Endocrinology & Metabolism, Amsterdam UMC, University of Amsterdam, Amsterdam, The Netherlands
[e] Endocrine Laboratory, Department of Clinical Chemistry, Amsterdam Gastroenterology, Endocrinology & Metabolism, Amsterdam UMC, Vrije Universiteit, Amsterdam, The Netherlands
[f] Department of Pediatrics, Division of Metabolic Disorders, Emma Children's Hospital, Amsterdam Gastroenterology, Endocrinology & Metabolism, Amsterdam UMC, University of Amsterdam, Amsterdam, The Netherlands
[g] Department of Computer Science, Vrije Universiteit, Amsterdam, The Netherlands
[h] Department of Clinical Chemistry, Amsterdam UMC, Vrije Universiteit & University of Amsterdam, Amsterdam Gastroenterology, Endocrinology & Metabolism, Amsterdam, The Netherlands

ARTICLE INFO

Keywords:
Congenital hypothyroidism
Neonatal screening
Machine learning
Random forest

ABSTRACT

Objective: The Dutch Congenital hypothyroidism (CH) Newborn Screening (NBS) algorithm for thyroidal and central congenital hypothyroidism (CH-T and CH-C, respectively) is primarily based on determination of thyroxine (T4) concentrations in dried blood spots, followed by thyroid-stimulating hormone (TSH) and thyroxine-binding globulin (TBG) measurements enabling detection of both CH-T and CH-C, with a positive predictive value (PPV) of 21%. A calculated T4/TBG ratio serves as an indirect measure for free T4. The aim of this study is to investigate whether machine learning techniques can help to improve the PPV of the algorithm without missing the positive cases that should have been detected with the current algorithm.

Design & methods: NBS data and parameters of CH patients and false-positive referrals in the period 2007–2017 and of a healthy reference population were included in the study. A random forest model was trained and tested using a stratified split and improved using synthetic minority oversampling technique (SMOTE). NBS data of 4668 newborns were included, containing 458 CH-T and 82 CH-C patients, 2332 false-positive referrals and 1670 healthy newborns.

Results: Variables determining identification of CH were (in order of importance) TSH, T4/TBG ratio, gestational age, TBG, T4 and age at NBS sampling. In a Receiver-Operating Characteristic (ROC) analysis on the test set, current sensitivity could be maintained, while increasing the PPV to 26%.

Conclusions: Machine learning techniques have the potential to improve the PPV of the Dutch CH NBS. However, improved detection of currently missed cases is only possible with new, better predictors of especially CH-C and a better registration and inclusion of these cases in future models.

Abbreviations: CH, congenital hypothyroidism; CH-C, central congenital hypothyroidism; CH-T, thyroidal congenital hypothyroidism; NBS, newborn screening; ROC, receiver operating characteristic; SMOTE, synthetic minority oversampling technique; PPV, positive predictive value; T4, thyroxine; TBG, thyroxine-binding globulin; TSH, thyroid stimulating hormone.

* Corresponding author at: Department of Clinical Chemistry, Amsterdam Gastroenterology, Endocrinology & Metabolism, Amsterdam UMC, University of Amsterdam, Meibergdreef 9, B.1-235, Amsterdam, North-Holland 1105 AZ, The Netherlands.
E-mail address: r.dejonge1@amsterdamumc.nl (R. de Jonge).

https://doi.org/10.1016/j.clinbiochem.2023.03.001
Received 6 November 2022; Received in revised form 28 February 2023; Accepted 2 March 2023
Available online 4 March 2023

1. Introduction

Congenital hypothyroidism (CH) is thyroid hormone deficiency present at birth, mostly caused by defective thyroid gland development or hormone biosynthesis (primary or thyroidal CH, CH-T). An equally important, less frequently occurring cause is hypothalamic or pituitary dysfunction resulting in secondary or central CH (CH-C) [1]. Early detection of CH by newborn screening (NBS) successfully contributes to adequate treatment and the prevention of neurodevelopmental disabilities, associated with this disorder [2–4].

Most NBS algorithms detect CH-T by a low total thyroxine (T4) blood concentration in combination with increased thyroid stimulating hormone (TSH) concentration. These TSH based programs will miss newborns with CH-C because they have low T4 concentrations in combination with normal or low serum TSH. Primary T4-based NBS algorithms thus potentially also detect newborns with CH-C, however unfortunately often at the cost of false-positive referrals due to low T4 caused by (non-thyroidal) illness or T4-binding globulin (TBG) deficiency [1,5–6]). In the Netherlands NBS program, T4 is measured in all newborns, followed by TSH measurement in case the T4 concentration is within the lowest 20% of the day. TBG is measured in the lowest 5% of daily T4 concentrations, and a calculated T4/TBG ratio is used to reduce false-positive referrals due to benign TBG deficiency [6–7]. Weight and (gestational) age are not used in the current NBS algorithm except in premature newborns with a birth weight of ≤2500 g and a gestational age of ≤36 (+0 days) weeks. Infants that received the heel puncture on or later than day 60 of life are also subject to the decision rule for premature newborns and referred based on only TSH (only measured in newborns with a T4 ≤ −0.8 Standard Deviation (SD)). The Dutch T4-TSH-TBG based algorithm effectively detected 86 CH-C patients in the period 2007–2017 but at the cost of a low positive predictive value (PPV) of 21%, mainly due to the screening criteria T4 and T4/TBG ratio [6]. In addition, the currently used T4-TSH-TBG algorithm does not have 100% sensitivity; it is generally accepted that a few CH-C cases will be missed and diagnosed later in childhood [13].

The aim of this study is to investigate whether machine learning can help to improve the Dutch CH screening algorithm by reducing the number of false-positive referrals (and increase the PPV) while maintaining the algorithm's current sensitivity [6].

2. Methods

A national database containing data from NBS referrals (newborns with CH as well as false-positive referrals) in the period 2007–2017 was used for this study [6]). The NBS algorithm for CH and the assays used, performance and cut offs were extensively described [6]. All CH-C and CH-T cases were included in this study; from the 86 CH-C cases, 4 were excluded because gestational age was missing. In short, between 72 and 168 h after birth, heel puncture blood is collected on filter paper and sent to one of the five regional screening laboratories in the Netherlands. Based on a National algorithm using 5 criteria [6], patients are referred to a clinician; inconclusive results are referred for a second heel puncture. **Criterion 1: T4 ≤ −3.0 SD & TBG >40 nmol/L blood**. Newborns with an abnormal TSH concentration are referred based on referral **criterion 2: T4 ≤ −0.8 SD & TSH ≥22 mIU/L blood**. A borderline TSH concentration (8≤ TSH <22 mIU/L blood) and/or a borderline T4 (−3.0< T4 ≤ −1.6 SD) in combination with a T4/TBG ratio ≤17 are considered inconclusive results, leading to a request for a second blood sample. In all second heel punctures T4, TSH and TBG are measured, and subsequently all newborns with a T4 ≤ −3.0 SD are immediately referred (referral **criterion 3: T4 ≤ −3.0 SD**). Newborns with a borderline (8≤ TSH <22 mIU/L blood) or abnormal TSH (≥22 mIU/L blood) concentration (referral **criterion 4: TSH ≥8 mIU/L blood**) and/or a borderline T4 (−3.0< T4 ≤ −1.6 SD) in combination with a T4/TBG ratio ≤17 (referral **criterion 5: −3.0 < T4 ≤ −1.6 SD & T4/TBG ratio ≤17**), are also referred. Premature newborns often have a low T4 concentration associated with prematurity or illness. Therefore, newborns with a birth weight ≤2500 g and a gestational age of ≤36 (+0 days) weeks are only assessed for their TSH, with the abovementioned cut-off values. This also applies to infants that received the heel puncture ≥ day 60 of life.

Referral rates are: 16.3% CH-T, 2.8% CH-C, 1% CH-unknown and 79.9% false-positive. Documented variables include gestational age, birth weight, age at NBS sampling, T4, TSH, TBG concentrations and the T4/TBG ratio. Approval was obtained from the NBS privacy committee of the Netherlands Organization for Applied Scientific Research (TNO, department of Child Health). This study complied with the World Medical Association Declaration of Helsinki regarding ethical conduct of research involving human subjects and/or animals. By nature of the Dutch screening algorithm [6], T4 is measured in all newborns (100%) followed by TSH measurement in the 20% lowest T4 of the day and TBG measurement in the 5% lowest T4 of the day. Therefore, in the National dataset, T4 is measured in 100%, TSH in 20% and TBG in 5% of newborns. Since the National database [6] did not include results of healthy newborns with a negative screening result, data (n = 1670) of a recent study to establish neonatal reference intervals for T4, TSH, TBG and the T4/TBG ratio were added to the data set [8]. For this dataset, T4, TSH and TBG were measured in 100% of healthy newborns. There were no missing features in both datasets and hence, no imputation was applied. The data set was split into a training (67%) and a test set (33%) using a stratified split, retaining the actual (unbalanced) ratio between healthy versus CH (N = 13%). To improve the training of the random forest [9] of an unbalanced dataset, we used synthetic minority oversampling technique (SMOTE) in the training set to create synthetic data points based on existing CH data [10]. For SMOTE the default settings were employed using traincontrol function in the Caret package (perc.over = 200, k = 5, perc.under = 200). A random forest model was trained and hyperparameters of the approach were tuned using 10-fold stratified cross-validation on the training set to select the optimal settings for these hyperparameters, being a random forest with 500 trees and 4 variables tried at each split. A random forest is a commonly used model in machine learning and creates a forest of trees. Each tree in the forest is trained on a randomly drawn sample of the training set and when building these trees a random subset of variables can only be used to increase diversity of trees. When the forest is used for classification, an aggregate score is taken over all trees. Diagnostic accuracy was calculated using receiver operating characteristic (ROC) analyses whereby the per class probabilities of the random forest classifier as generated using the Caret package are used to create the ROC curve. Different points on the ROC curve are then chosen based on sensitivity (resulting in a threshold selection of the random forest), whereby the PPV is calculated using the selected threshold. Variable importance was determined by decrease in accuracy when removing variables from the model using the Gini Index. This is done for all trees and the difference between the two accuracies are then averaged over all trees and normalized by the standard error. Because we do not know how many cases are missed in the current NBS algorithm, 100% sensitivity for the machine learning algorithm was defined as detection of all positive cases detected by the current NBS algorithm. The aim of this study was to improve the PPV of the current NBS algorithm and hence, we could compare the PPV of the current algorithm [6] with the new machine learning algorithm because it was applied to the same data set with the same CH prevalence. Based on previous work, registries and clinical experience, no patients with CH-T and approximately 2–4 patients/year with CH-C are missed by the current NBS algorithm and hence, the reported CH prevalence in this study is slightly underestimated. All analyses were performed in R (version 3.6.3) and Caret (version 6) [11–12].

3. Results

Data of 4668 newborns (gestational age (mean ± SD): 39.5±1.4 weeks) were included in this study; 458 CH-T and 82 CH-C patients,

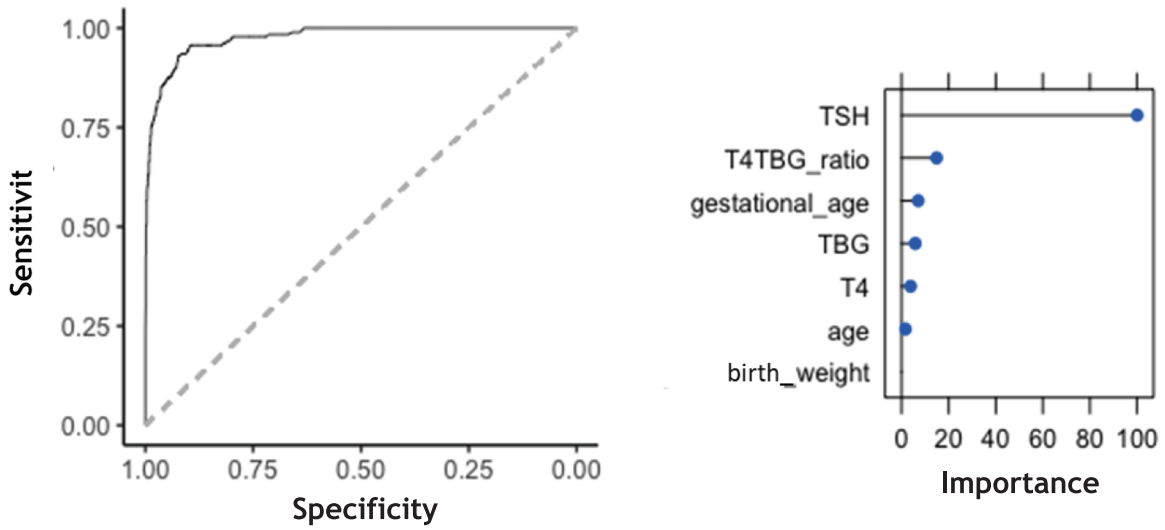

Fig. 1. Random forest model of CH vs. healthy population: ROC analysis (ratios) and importance plot (%) on the test set.

Table 1
Diagnostic accuracy of the random forest model under different thresholds based on selected sensitivity.

Specificity	Sensitivity	Accuracy	TN	TP	FN	FP	PPV
0.63	1.0	0.68	890	184	0	514	0.26
0.66	0.99	0.70	929	182	2	475	0.28
0.71	0.98	0.74	1000	181	3	404	0.31
0.74	0.98	0.77	1037	180	4	367	0.33

2332 false-positive referrals [6] and 1670 healthy newborns. In order of importance, variables determining identification of CH were TSH, T4/TBG ratio, gestational age, TBG, T4 and age at NBS sampling (also referred to as 'age'). Birth weight was not of importance for the model (Fig. 1) probably because it is correlated with gestational age. In a ROC analysis on the test set, a sensitivity of 100% could be maintained, while reaching a model specificity of 63% and a PPV of 26% (Table 1). Other cut-offs on the ROC could be chosen, allowing a sensitivity of 98–99% with a specificity of 66–74% and PPV of 28–33%. Possible sensitivity, specificity, accuracy and PPV of the model with the corresponding numbers of true negatives, false-negatives, true-positives, and false-positives are reported in the table below the ROC (Fig. 1, Table 1). Results with a sensitivity <98% resulted in more than 4 missed cases in the test set and were considered undesired in the context of this study. SMOTE improved diagnostic accuracy, the area under the ROC increased from 0.85 (95% CI: 0.76–0.93) to 0.89 (95% CI: 0.87–0.90; test set).

CH: congenital hypothyroidism, FN: false-negatives, FP: false-positives, PPV: positive predictive value, ROC: receiver operator characteristic, TN: true-negatives, TP: true-positives.

4. Discussion

We show that a random forest machine learning algorithm improved the current PPV of the Dutch NBS algorithm from 21% [6] to 26% while maintaining 100% sensitivity. Increasing the PPV to 26% using our Machine Learning algorithm would reduce the number of false-positive referrals by 48 per year. Lowering the desired 100% sensitivity to a sensitivity of 99 or 98%, increased the PPV to 28–33%.

The low PPV of the current Dutch NBS algorithm for CH is mainly caused by the criteria used to also detect CH-C cases [13]. We confirmed that the T4/TBG ratio was the most important variable in a separate random forest model for CH-C vs. healthy children with a PPV of 5% (results not shown), in line with the performance of the current NBS algorithm [6]. Using decision trees, we confirmed the appropriateness of cut-off values in our current NBS algorithm (data not shown). Hence, other CH-C predictors should be investigated and added to the currently used T4 and TBG to improve the PPV in CH-C NBS. In a recent study using a random forest model to improve the PPV of NBS for metabolic disorders, metabolic screening analytes were far more important than generic co-variates such as gestational age and weight [14]. We had similar findings with a major importance of parameters TSH and the T4/TBG ratio, alongside minor importance of gestational age.

The disadvantage of our approach is that we used a new technique on an incomplete data set of an existing algorithm (TSH and TBG in the lowest 20% and 5 % T4 of the day, respectively) [6]. Despite these limitations we could improve the PPV by 5% without missing CH cases. Ideally, we would like to start a prospective study measuring all variables on all dried blood spots and add new features. This will allow development of a completely new screening algorithm with improved diagnostic accuracy in detecting CH-C patients maintaining the current accuracy for CH-T patients and reducing false-positive referrals caused by detecting CH-C patients. Our study demonstrates that machine learning algorithms have the potential to improve newborn screening algorithms in line with other studies [14].

Changes in assay performance over time [6] may influence a machine learning model performance. In general, improved assay standardization will result in better diagnostic accuracy for NBS screening algorithms. However, the random forests approach is robust against variation in the data because trees are built in different samples of the data and different subsets of features, thus the larger variation in TSH and small change in TBG variation over time [6] probably did not have much influence on the model.

The reported accuracy of the model should be interpreted with caution, as it does not reflect the actual performance of the model on the complete Dutch newborn population. The random forest model included only a small fraction of healthy newborns (screen negatives), as compared to reality, therefore skewing the results. Also, it must be noted that the reported 100% sensitivity of the model reflects the population of patients that is presented to the model. Unfortunately, the current Dutch NBS for CH does not have a 100% sensitivity, and although detection of CH-C by the Dutch approach is better than in T4-based NBS strategies without TBG measurement [6–7], CH-C cases are still missed and diagnosed later in childhood [15]. Reducing the number of missed cases will only be possible if data on these cases are included in the model. Unfortunately, complete registration of these cases is currently lacking.

In conclusion, machine learning techniques have the potential to improve the PPV of the Dutch CH NBS but could not improve the diagnostic accuracy for detecting specific CH-C cases using the current features of the NBS program. Before replacing the current algorithm by machine learning techniques, the model will need to run in parallel to the current NBS for several years and be re-evaluated when more data is available. Future research would ideally include data of all current screening parameters and (yet unknown) predictors of CH in all newborns and missed patients, requiring a large prospective study and complete registration of missed cases, possibly also constructing separate algorithms to detect CH-T and CH-C.

Declaration of Competing Interest

Annet M. Bosch has received a speaker's fee from Nutricia and has been a member of advisory boards for Biomarin.

References

[1] M.V. Rastogi, S.H. LaFranchi, Congenital hypothyroidism, Orphanet J Rare Dis. 5 (2010) 17, https://doi.org/10.1186/1750-1172-5-17.

[2] G. Ford, S.H. LaFranchi, Screening for congenital hypothyroidism: a worldwide view of strategies, Best Pract Res Clin Endocrinol Metab. 28 (2014) 175–187, https://doi.org/10.1016/j.beem.2013.05.008.

[3] J. Simoneau-Roy, S. Marti, C. Deal, C. Huot, P. Robaey, G. Van Vliet, Cognition and behavior at school entry in children with congenital hypothyroidism treated early with high-dose levothyroxine, J Pediatr. 144 (2004) 747–752, https://doi.org/10.1016/j.jpeds.2004.02.021.

[4] W.F. Simons, P.W. Fuggle, D.B. Grant, I. Smith, Intellectual development at 10 years in early treated congenital hypothyroidism, Archives of Disease in Childhood. 71 (3) (1994) 232–234.

[5] M.J. Kilberg, I.R. Rasooly, S.H. LaFranchi, A.J. Bauer, C.P. Hawkes, Newborn Screening in the US May Miss Mild Persistent Hypothyroidism, J Pediatr. 192 (2018) 204–228, https://doi.org/10.1016/j.jpeds.2017.09.003.

[6] K. Stroek, A.C. Heijboer, M.J. Bouva, C.P.B. van der Ploeg, M.A. Heijnen, G. Weijman, A.M. Bosch, R. de Jonge, P.C.J.I. Schielen, A.S.P. van Trotsenburg, A. Boelen, Critical evaluation of the newborn screening for congenital hypothyroidism in the Netherlands, EJE 19 (2020) 1048, https://doi.org/10.1530/eje-19-1048.

[7] C.I. Lanting, D.A. van Tijn, J.G. Loeber, T. Vulsma, J.J. de Vijlder, P.H. Verkerk, Clinical effectiveness and cost-effectiveness of the use of the thyroxine/thyroxine-binding globulin ratio to detect congenital hypothyroidism of thyroidal and central origin in a neonatal screening program, Pediatrics 116 (2005) 168–173, https://doi.org/10.1542/peds.2004-2162.

[8] K. Stroek, A.C. Heijboer, M. van Veen-Sijne, A.M. Bosch, C.P.B. van der Ploeg, N. Zwaveling-Soonawala, R. de Jonge, A.S.P. van Trotsenburg, A. Boelen, Improving the Dutch Newborn Screening for Central Congenital Hypothyroidism by Using 95% Reference Intervals for Thyroxine-Binding Globulin, Eur Thyroid J 10 (3) (2021) 222–229.

[9] L. Breiman, Random forests. Machine learning 45 (2002) 5–32.

[10] N. Chawla, K. Bowyer, L. Hall, W. Kegelmeyer, SMOTE: Synthetic Minority Oversampling Technique, J. Artif. Intell. Res. (JAIR) 16 (2020) 321–357, https://doi.org/10.1613/jair.953.

[11] Kuhn M. 2022. caret: Classification and Regression Training. R package version 6.0-93. URL https://eur04.safelinks.protection.outlook.com/?url=https%3A%2F%2Fcran.r-project.org%2Fpackage%3Dcaret&data=05%7C01%7Ca.s.vantrotsenburg%40amsterdamumc.nl%7Cb487f850a09b4e0bea5108db081729d2%7C68dfab1a11bb4cc6beb528d756984fb6%7C0%7C0%7C638112667798736625%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=rvii8PHtiiSJFRw9fT5qL7lqvCfj0sovu4ZLmPjFYdo%3D&reserved=0.

[12] R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://eur04.safelinks.protection.outlook.com/?url=https%3A%2F%2Fwww.r-project.org%2F&data=05%7C01%7Ca.s.vantrotsenburg%40amsterdamumc.nl%7Cb487f850a09b4e0bea5108db081729d2%7C68dfab1a11bb4cc6beb528d756984fb6%7C0%7C0%7C638112667798892852%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=ekVQe81Fg5Hu4F7P66LD0AwhriHm33%2BsYpJ8Dul%2FdKM%3D&reserved=0.

[13] D.S. Saleh, S. Lawrence, M.T. Geraghty, P.H. Gallego, K. McAssey, D.K. Wherrett, P. Chakraborty, Prediction of congenital hypothyroidism based on initial screening thyroid-stimulating-hormone, BMC pediatrics. 16 (2016) 24, https://doi.org/10.1186/s12887-016-0559-0.

[14] G. Peng, Y. Tang, T.M. Cowan, G.M. Enns, H. Zhao, C. Scharfe, Reducing False-Positive Results in Newborn Screening Using Machine Learning, International journal of neonatal screening 6 (2020) 16, https://doi.org/10.3390/ijns6010016.

[15] L. van Iersel, H.M. van Santen, G.R.J. Zandwijken, N. Zwaveling-Soonawala, A.C. S. Hokken-Koelega, A.S.P. van Trotsenburg, Low FT4 Concentrations around the Start of Recombinant Human Growth Hormone Treatment: Predictor of Congenital Structural Hypothalamic-Pituitary Abnormalities? Horm Res Paediatr. 89 (2018) 98–107, https://doi.org/10.1159/000486033.