Pregnancy Hypertension: An International Journal of Women's Cardiovascular Health 6 (2016) 53–59

Contents lists available at ScienceDirect

Pregnancy Hypertension: An International Journal of Women's Cardiovascular Health

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

Extending the scope of pooled analyses of individual patient biomarker data from heterogeneous laboratory platforms and cohorts using merging algorithms

Órlaith Burke [a,*,1,2,3], Samantha Benton [b,1,2,3], Pawel Szafranski [c,1], Peter von Dadelszen [b,1], S. Catalin Buhimschi [d], Irene Cetin [e], Lucy Chappell [f], Francesc Figueras [g], Alberto Galindo [h], Ignacio Herraiz [h], Claudia Holzman [i], Carl Hubel [j], Ulla Knudsen [k], Camilla Kronborg [l], Hannele Laivuori [m], Olav Lapaire [n], Thomas McElrath [o], Manfred Moertl [p], Jenny Myers [q], Roberta B. Ness [r], Leandro Oliveira [s], Gayle Olson [t], Lucilla Poston [f], Carrie Ris-Stalpers [u], James M. Roberts [j], Sarah Schalekamp-Timmermans [v], Dietmar Schlembach [w], Eric Steegers [v], Holger Stepan [x], Vassilis Tsatsaris [y], Joris A. van der Post [u], Stefan Verlohren [z], Pia M. Villa [aa], David Williams [bb], Harald Zeisler [cc], Christopher W.G. Redman [c,1], Anne Cathrine Staff [dd,1], for the Global Pregnancy Collaboration

[a] *Nuffield Department of Population Health, University of Oxford, Oxford, UK*
[b] *Department of Obstetrics and Gynaecology and Child and Family Research Institute, University of British Columbia, Vancouver, British Columbia, Canada*
[c] *Nuffield Department of Obstetrics and Gynaecology, University of Oxford, Oxford, UK*
[d] *Yale University, New Haven, CT, USA*
[e] *University of Milan, Italy*
[f] *Women's Health Academic Centre, King's College London, London, UK*
[g] *Hospital Clinic, University of Barcelona, Barcelona, Spain*
[h] *Fetal Medicine Unit, Dept of Obstetrics and Gynecology, Hospital Universitario 12 de Octubre and Universidad Complutense, Madrid, Spain*
[i] *Michigan State University, East Lansing, MI, USA*
[j] *Magee-Womens Research Institute, University of Pittsburgh, Pittsburgh, USA*
[k] *Dept. of Obstetrics and Gynecology, Aarhus University Hospital and Aarhus University, Aarhus, Denmark*
[l] *Dept. of Oncology, Aarhus University Hospital and Aarhus University, Aarhus, Denmark*
[m] *Medical Genetics, Obstetrics and Gynaecology and Institute for Molecular Medicine Finland, University of Helsinki and Helsinki University Hospital, Helsinki, Finland*
[n] *University of Basel, Basel, Switzerland*
[o] *Brigham & Women's Hospital, Harvard Medical School, Boston, MA, USA*
[p] *Interdisciplinary Centre for Gynecology, Gyneco-oncology and Feto-Maternal Medicine, Klagenfurt & Medical University, Graz, Austria*
[q] *Maternal and Fetal Health Research Centre, St. Mary's Hospital and University of Manchester, Manchester, UK*
[r] *University of Texas, School of Public Health, Houston, TX, USA*
[s] *Obstetrics Department, Universidade Estadual Paulista, Botucatu, Brazil*
[t] *University of Texas Medical Branch, Galveston, TX, USA*
[u] *Department of Obstetrics and Gynaecology, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands*
[v] *Erasmus Medical Centre, Rotterdam, The Netherlands*
[w] *University of Jena, Germany*
[x] *University of Leipzig, Germany*
[y] *Université Paris Descartes, Paris, France*
[z] *Charité University Medicine, Berlin, Germany*
[aa] *Obstetrics and Gynecology, University of Helsinki and Helsinki University Hospital, Helsinki, Finland*
[bb] *Institute for Women's Health, NIHR University College London Hospitals Biomedical Research Centre, London, UK*
[cc] *Department of Obstetrics and Gynecology, Medical University Vienna, Austria*
[dd] *Department of Obstetrics and Gynecology, Oslo University Hospital and University of Oslo, Oslo, Norway*

Abbreviations: PlGF, placental growth factor; HDP, hypertensive disorders of pregnancy; BRC, best reference curve; IPD, individual patient data; GA, gestational age; MoM, Multiple of the Median; CI, confidence interval.

* Corresponding author at: Nuffield Department of Population Health, University of Oxford, Old Road Campus, Headington, Oxford OX37LF, United Kingdom.
E-mail address: orlaith.burke@ndph.ox.ac.uk (Ó. Burke).

[1] Members of Global Collaboration Group.
[2] Members of Global Collaboration Angiogenic Factor Protocol Committee.
[3] Joint first authors.

http://dx.doi.org/10.1016/j.preghy.2015.12.002

ARTICLE INFO

Article history:
Received 16 October 2015
Received in revised form 19 November 2015
Accepted 8 December 2015
Available online 12 January 2016

Keywords:
Merging algorithms
Individual patient data
Best reference curve
Pooled analysis
Biomarker data
Pre-eclampsia

ABSTRACT

Background: A common challenge in medicine, exemplified in the analysis of biomarker data, is that large studies are needed for sufficient statistical power. Often, this may only be achievable by aggregating multiple cohorts. However, different studies may use disparate platforms for laboratory analysis, which can hinder merging.
Methods: Using circulating placental growth factor (PlGF), a potential biomarker for hypertensive disorders of pregnancy (HDP) such as preeclampsia, as an example, we investigated how such issues can be overcome by inter-platform standardization and merging algorithms. We studied 16,462 pregnancies from 22 study cohorts. PlGF measurements (gestational age $\geqslant$20 weeks) analyzed on one of four platforms: R&D® Systems, Alere®Triage, Roche®Elecsys or Abbott®Architect, were available for 13,429 women. Two merging algorithms, using Z-Score and Multiple of Median transformations, were applied.
Results: Best reference curves (BRC), based on merged, transformed PlGF measurements in uncomplicated pregnancy across six gestational age groups, were estimated. Identification of HDP by these PlGF-BRCs was compared to that of platform-specific curves.
Conclusions: We demonstrate the feasibility of merging PlGF concentrations from different analytical platforms. Overall BRC identification of HDP performed at least as well as platform-specific curves. Our method can be extended to any set of biomarkers obtained from different laboratory platforms in any field. Merged biomarker data from multiple studies will improve statistical power and enlarge our understanding of the pathophysiology and management of medical syndromes.

1. Introduction

Large datasets are essential for sufficient statistical power to characterize subsets of disease. The usefulness of single cohorts can be enhanced by combining several studies to facilitate analyses of pooled individual patient data (IPD). However, to date such studies have only collected primary outcomes measured on comparable scales or, in the case of biomarkers, using the same assay platforms. Different assay platforms may vary in their sensitivity, precision, and concentration ranges. In such cases, valid methods of standardization of laboratory data are required in order to aggregate individual patient data.

The Global Pregnancy Collaboration (CoLAB) was established in 2011 (http://pre-empt.cfri.ca/colaboratory/global-pregnancy-collaboration) to facilitate data and sample sharing between research groups studying preeclampsia and other pregnancy disorders. Preeclampsia is a hypertensive disorder of pregnancy which complicates 3–4% of pregnancies and is a leading cause of maternal and fetal/neonatal mortality and morbidity worldwide. Because preeclampsia is clinically and biologically heterogeneous, (e.g. early and late disease have different prognoses and perhaps etiologies) improvements in management, prediction, diagnosis, prevention and treatment have been difficult to achieve [1–3].

Circulating maternal biomarkers of placental origin have been proposed as novel tools for identifying hypertensive disorders of pregnancy (HDP). However, to date, precise estimates of diagnostic sensitivity and specificity have yet to be achieved because individual studies have been too small. Clinical data can be easily standardized for aggregation of cohorts, but laboratory biomarker data present the unique problem that they often use different analytical platforms with different ranges and results.

This paper focuses on clinical and laboratory data for placenta protein, placental growth factor (PlGF) to predict and/or diagnose hypertensive disorders of pregnancy (HDP). These disorders are associated with severe reductions in circulating PlGF concentrations [1,4,5]. In the cohorts included in this study, PlGF was quantified on one of four laboratory platforms, each with different analytic performance. We developed a method of standardizing PlGF data to allow pooling. Additionally, concentrations of PlGF are known to change with gestational age (GA) and to show the power of the pooled data, we developed a best reference curve (BRC) over six gestational age groups. The rate of accurate identification of women with HDP using the merged BRC was compared to unmerged (platform-specific) rates.

This paper demonstrates a principle that can be generalized to the study of other biomarkers for any complex, heterogeneous medical conditions requiring large cohorts to draw useful conclusions, which also use different assay platforms to measure the same biomarker.

2. Materials and methods

2.1. Study database

In 2011–2012, we invited principal investigators with studies of circulating maternal angiogenic factors in pregnancy to participate in this study. We included any study in which maternal blood samples were collected at least once at any time during pregnancy (uncomplicated or otherwise) and had been analyzed for PlGF. Adequate clinical, demographic and pregnancy outcome information was necessary for inclusion. 22 cohorts were included in the present analyses (Supplementary material Table 1, with references to detailed information about each study, including individual patient consent and formal study research ethical approval). The datasets varied in sample size, maternal demographics as well as study design, including both low and high risk pregnancies. Missing data were retrieved, where possible. Individual datasets were integrated into one central database, which was cleaned and checked to ensure data integrity was maintained. Reported measures of PlGF below the limit of detection for each of the four platforms were recorded as the threshold value. These occurred in less than 1.5% of the observations and were not removed because these observations are expected to include the most severe cases of placental dysfunction associated with HDP.

The final database contained information on 16,462 pregnancies. Here we included only those women (n = 13,429) who had at least one PlGF measurement at or after 20 weeks' gestation (the time when preeclampsia presents clinically, by definition). Four different analytical platforms had been used by the included cohorts: Alere Triage PlGF, Roche Elecsys PlGF, R&D Systems PlGF and Beckman-Coulter PlGF. The number of pregnancies by cohort and analytical platform is listed in Supplementary material Table 1.

2.2. Flow-chart for the methodology

Supplementary material Fig. 1 is a flow-chart of the generalized methodology of this paper. Blue boxes outline the steps from definition of non-cases through to merging of data and estimation of the BRC. Red boxes highlight validation steps associated with particular elements of the methodology. Green boxes indicate additional information for the user.

2.3. Data transformations for normal pregnancies

The primary aim of the analysis was to merge PlGF measurements from the four platforms used by the 22 study cohorts. This was achieved by first considering the least variable group of observations (termed 'non-cases'): in our example this group comprised those women who had uncomplicated pregnancies. A non-case was defined here as any woman who delivered a live born infant at term ($\geqslant$37 weeks gestation) with a birthweight >10th percentile for gestational age at delivery and sex, and without HDP, fetal growth restriction, or gestational diabetes. Women with pre-existing hypertension and/or pre-existing diabetes were excluded.

7600 (56.6%) pregnancies met these non-case criteria. To ensure independence, each pregnancy contributed only one PlGF measurement (from blood drawn at or after 20 weeks' gestation) to the analysis. In pregnancies where multiple samples have been taken the measurement was randomly selected. The number of non-case observations for each platform is presented in Table 1 for the gestational age categories: 20–23^{+6}, 24–26^{+6}, 27–32^{+6}, 33–36^{+6}, 37–39^{+6}, 40+ weeks (where 20–23^{+6} includes measurements taken from the start of 20 weeks gestation to 23 weeks and 6 days gestation).

To determine whether non-case measurements could be standardized for subsequent merging across platforms, we considered two merging algorithms based on Z-Score and Multiples of the Median (MoM) transformations.

2.3.1. The Z-Score transformation

The Z-Score transformation assumes normality of non-case measurements within each subgroup. In our PlGF data, the measurements within each GA-platform subgroup were shown to be log-normally distributed. Hence, a log-transformation was needed to achieve normality. The estimated mean and standard deviation of the log-transformed PlGF measurements of each non-case platform-GA subgroup were used to transform non-case PlGF measurements as

$$y_i^Z = \frac{\ln(y_i) - \mu_{p[i]g[i]}}{\sigma_{p[i]g[i]}} \quad (1)$$

where y_i and y_i^Z are the original and Z-Score-transformed PlGF measurements respectively, $\mu_{p[i]g[i]}$ and $\sigma_{p[i]g[i]}$ correspond to the mean and standard deviation (of the log-transformed non-case subgroup) for the platform associated with the i^{th} patient ($p[i]$) and GA category associated with the i^{th} patient ($g[i]$).

If the assumption of normality holds then, by definition, each transformed platform-GA subgroup follows a standard Normal distribution (zero mean, unit standard deviation). Hence transformed data from all platforms may be merged within GA categories. If this assumption is not satisfied, the transformation is still valid as a method of standardization of data for merging (since the transformation provides standardized observations irrespective of being standard Normal).

2.3.2. The MoM transformation

The Multiple of the Median (MoM) transformation only requires the estimation of one parameter (m) to transform the measurements (y_i) in each platform-GA subgroup to a common scale:

$$y_i^{\text{MoM}} = \frac{\ln(y_i)}{m_{p[i]g[i]}} \quad (2)$$

where y_i and y_i^{MoM} are the original and MoM-transformed PlGF measurements and $m_{p[i]g[i]}$ denotes the sample median of the log-transformed platform-GA subgroup associated with the i^{th} patient. The MoM-transformed measurements within each GA category are on the same scale and may be merged.

2.4. Merging of transformed non-case measurements

Bootstrapping was used to estimate the parameters of both transformations for each platform-GA subgroup. The bootstrap estimates provide insight into the distribution of the parameter estimates, e.g., variability/precision of the estimates. The transformations were then applied to individual PlGF measurements for all pregnancies defined as non-cases. These transformed PlGF measurements across all four platforms were merged within each GA category.

2.5. Validation of merging

Merged data plots (not shown) were used to assess the degree of merging within each GA category. K-means clustering analysis was used to determine whether distinct groups within the merged data set were identifiable.

2.6. Validation of parameter estimate

Leave-one-out cross-validation (LOO-CV) was used to measure the possible influence of each cohort on parameter estimation for both transformations. Bootstrap empirical confidence intervals were used to determine the significance of cohort effects. No single cohort had an effect on the estimation of a single parameter in all GA categories, again supporting a valid merging process of our 22 heterogeneous cohorts.

2.7. Reference curve thresholds and application

We extended the analysis to estimate the best reference curve (BRC) for transformed PlGF concentrations over gestational age. The merged data in each GA category were used to estimate a reference curve. Thresholds at the 5th percentile (along with the associated 95% empirical confidence interval) were estimated empirically using bootstrap samples of the transformed non-case PlGF measurements in each GA category.

We applied the BRC to the identification of pregnancies complicated by hypertensive disorders as an example here. We compared the performance of the merged 5th percentile thresholds to that of the corresponding platform-specific thresholds, in identifying pregnancies with any HDP outcome (termed a "case").

For illustrative purposes only, our *case definition* was any woman who had a final diagnosis of gestational hypertension, preeclampsia, super-imposed preeclampsia, HELLP syndrome (a form of severe preeclampsia comprising hemolysis, elevated liver enzymes and low platelets) or eclampsia occurring after 20 weeks gestation. Gestational hypertension and preeclampsia were defined by the individual cohorts according to the conventionally used definition; new onset-hypertension ($\geqslant$140/90) $\geqslant$GA 20 weeks, together with new-onset proteinuria in the case of preeclampsia. Only pregnancies with PlGF measurements from blood sampled at the time of diagnosis or within 2 weeks prior to diagnosis were included. As before, each woman only contributed a single measurement to the analysis. There were 1423 pregnancies meeting these criteria (Table 1).

Table 1
Non-case and case sample sizes by gestational age category and platform.

Platform GA category	Non-cases					Cases				
	R&D	Alere	Roche	Abbott	GA category total	R&D	Alere	Roche	Abbott	GA category total
20–23^{+6}	434	117	88	3900	4539	1	2	7	2	12
24–26^{+6}	266	35	144	395	840	19	22	28	11	80
27–32^{+6}	78	462	152	178	870	64	122	78	26	290
33–36^{+6}	69	171	91	394	725	96	140	156	95	487
37–39^{+6}	76	96	66	145	383	67	147	80	133	427
40+	6	99	41	97	243	18	36	24	49	127
Platform Total	929	980	582	5109	7600	265	469	373	316	1423

PlGF measurements from cases were transformed using the Z-Score and MoM algorithms (as described above) for comparison with the BRC. LOO-CV was used to measure both the possible cohort influences and possible platform influences on parameter estimation. Performance was evaluated by the rate of correct identification of cases for the merged and platform-specific thresholds.

3. Results

3.1. Z-Score transformation parameter estimates

The mean ($\mu_{p[i]g[i]}$) and standard deviation ($\sigma_{p[i]g[i]}$) for the Z-Score transformation of the log-transformed PlGF measurements were estimated using 10,000 bootstrap iterations. The estimated mean and standard deviation of the non-case PlGF measurements in each platform-GA subgroup are presented in Table 2 alongside the associated bootstrap standard errors. These estimates were used to transform the original PlGF measurements (Eq. (1)). The transformed datasets were tested for normality using the Anderson–Darling test (data not shown). For 62.5% of the platform-GA subgroups (15 of 24) the null hypothesis (that the distributions are Normal) was not rejected at the 5% significance level (p-values ranging from 0.06 to 0.91).

Kolmogorov–Smirnov tests (data not shown) confirmed that in each GA category, transformed non-case measurements on each of the four platforms followed the same distribution, thus validating our decision to merge these transformed data sets.

3.2. MoM transformation parameter estimates

The median of the log-transformed PlGF measurements in each platform-GA group ($m_{p[i]g[i]}$) to be used in the Multiple of the Median transformation algorithm was estimated with 10,000 bootstrap iterations (Table 2). Kolmogorov–Smirnov tests (data not shown) confirmed the suitability of merging MoM transformed measurements from the four platforms within each GA category. These estimated parameters were used in the MoM transformation of the original PlGF measurements (Eq. (2)).

The PlGF measurements on all four platforms are measured in pictograms per milliliter (pg/ml). The transformed PlGF measurements do not have equivalent physical units. The transformed measurements are therefore referred to as the PlGF Z-Scores (or PlGF MoMs).

3.3. Validation of merging

Merging was deemed successful as clustering analysis could not identify platform-specific groups of merged PlGF Z-Scores (or PlGF MoMs) in any GA category. To further compare these algorithms, the Adjusted Rand Index was calculated for the k-means clustering technique ($k = 4$) in each GA category (Z-score range: −0.002, 0.004, MoM range: −0.011, 0.011) and showed that each algorithm provided a merged data set in which platform-specific groups were unidentifiable, justifying their merging (and later their use in creating a common best reference curve for PlGF).

Table 2
Estimated mean (μ) and standard deviation (σ) parameters for the Z-Score transformation and median (m) parameter for the MoM transformation of each platform-GA subgroup (*and Associated Bootstrap Standard Errors*).

Platform $p[i]$ GA $g[i]$		R&D		Alere		Roche		Abbott	
20–23^{+6}	μ	5.953	(0.029)	5.220	(0.068)	5.642	(0.049)	5.430	(0.008)
	σ	0.598	(0.020)	0.733	(0.063)	0.461	(0.030)	0.524	(0.007)
	m	5.945	(0.044)	5.289	(0.072)	5.596	(0.075)	5.405	(0.011)
24–26^{+6}	μ	6.345	(0.034)	5.885	(0.097)	6.021	(0.046)	6.067	(0.030)
	σ	0.559	(0.024)	0.569	(0.059)	0.552	(0.030)	0.602	(0.026)
	m	6.343	(0.043)	5.940	(0.133)	6.046	(0.106)	6.050	(0.030)
27–32^{+6}	μ	6.353	(0.075)	6.068	(0.036)	6.162	(0.056)	6.383	(0.052)
	σ	0.664	(0.053)	0.782	(0.031)	0.689	(0.036)	0.695	(0.039)
	m	6.465	(0.102)	6.095	(0.051)	6.184	(0.062)	6.378	(0.038)
33–36^{+6}	μ	5.673	(0.104)	5.462	(0.089)	5.836	(0.081)	5.938	(0.049)
	σ	0.855	(0.068)	1.172	(0.063)	0.766	(0.051)	0.985	(0.034)
	m	5.650	(0.160)	5.608	(0.084)	5.801	(0.142)	6.022	(0.043)
37–39^{+6}	μ	5.184	(0.074)	4.512	(0.119)	5.389	(0.098)	4.994	(0.742)
	σ	0.645	(0.067)	1.140	(0.063)	0.787	(0.066)	0.905	(0.043)
	m	5.076	(0.071)	4.470	(0.221)	5.387	(0.096)	4.957	(0.096)
40+	μ	5.258	(0.299)	4.010	(0.098)	4.918	(0.110)	4.796	(0.104)
	σ	0.685	(0.233)	0.974	(0.061)	0.705	(0.058)	1.026	(0.054)
	m	5.214	(0.329)	3.846	(0.169)	4.956	(0.178)	4.796	(0.127)

Table 3
Thresholds [and lower bound of associated 95% CI] estimated from merged and platform-specific transformed data.

GA category		$20–23^{+6}$		$24–26^{+6}$		$27–32^{+6}$		$33–36^{+6}$		$37–39^{+6}$		40+	
Transformation													
Z-Score	Merged	−1.53	[−1.48]	−1.66	[−1.51]	−1.73	[−1.63]	−1.77	[−1.62]	−1.56	[−1.47]	−1.54	[−1.36]
	R&D	−1.65	[−1.47]	−1.65	[−1.42]	−1.89	[−1.28]	−1.49	[−1.26]	−1.33	[−0.94]	−1.05	[−0.21]
	Alere	−1.65	[−1.28]	−1.54	[−1.11]	−1.74	[−1.60]	−1.97	[−1.59]	−1.67	[−1.46]	−1.30	[−1.12]
	Roche	−1.42	[−1.24]	−1.57	[−1.34]	−1.72	[−1.46]	−1.60	[−1.34]	−1.52	[−1.31]	−1.48	[−1.44]
	Abbott	−1.52	[−1.48]	−1.66	[−1.46]	−1.64	[−1.48]	−1.76	[−1.60]	−1.58	[−1.29]	−1.54	[−1.28]
MoM	Merged	0.85	[0.86]	0.84	[0.86]	0.79	[0.81]	0.69	[0.72]	0.72	[0.74]	0.71	[0.74]
	R&D	0.84	[0.85]	0.85	[0.87]	0.79	[0.85]	0.78	[0.81]	0.85	[0.90]	0.87	[0.98]
	Alere	0.76	[0.81]	0.85	[0.88]	0.77	[0.79]	0.56	[0.64]	0.58	[0.64]	0.74	[0.78]
	Roche	0.89	[0.91]	0.85	[0.88]	0.80	[0.83]	0.79	[0.83]	0.78	[0.81]	0.77	[0.79]
	Abbott	0.86	[0.86]	0.84	[0.86]	0.82	[0.84]	0.70	[0.72]	0.74	[0.72]	0.67	[0.72]

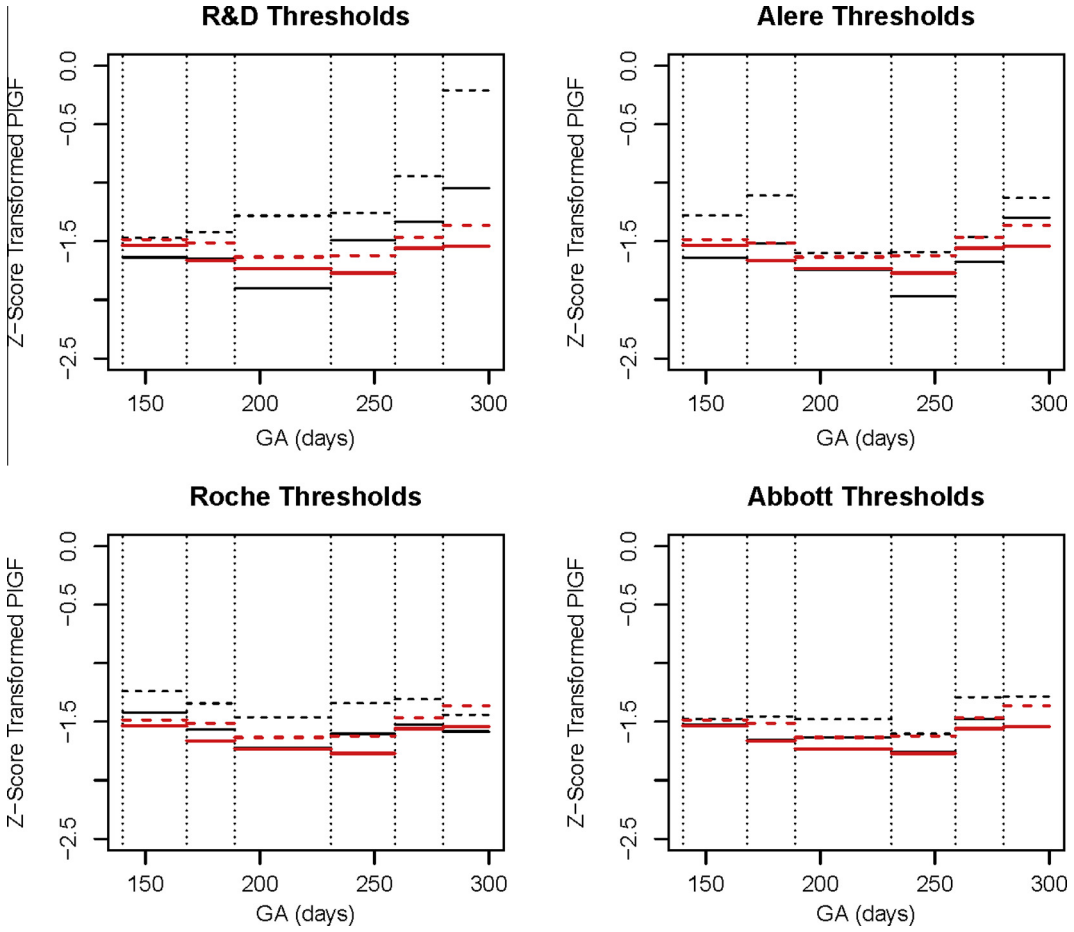

Fig. 1. Platform-specific thresholds (black) and the merged thresholds (red) for the Z-Score transformation.

3.4. Validation of parameter estimate

Leave-one-out cross-validation found that no single cohort had an effect on the estimation of a single parameter in all GA categories, again supporting a valid merging process of our 22 heterogeneous cohorts.

3.5. Reference curve thresholds and application

These merged thresholds are presented in Table 3. Note that the threshold values for PlGF Z-Scores and PlGF MoMs are not directly comparable since they are on different scales. LOO-CV demonstrated that no single cohort or platform had a significant effect on these estimates.

Similarly, thresholds (and associated 95% empirical confidence intervals) for the platform-GA subgroups of transformed PlGF measurements were constructed (Table 3). Fig. 1 illustrates the difference between the platform-specific thresholds and the merged thresholds for the Z-Score transformation. The platform-specific thresholds (shown in black) display much higher variability (shown by the wide spread between the mean threshold and the lower bound of its associated 95% empirical confidence interval) than the merged thresholds (shown in red).

A threshold was deemed to correctly identify a case if its transformed PlGF was lower than the estimated GA-specific

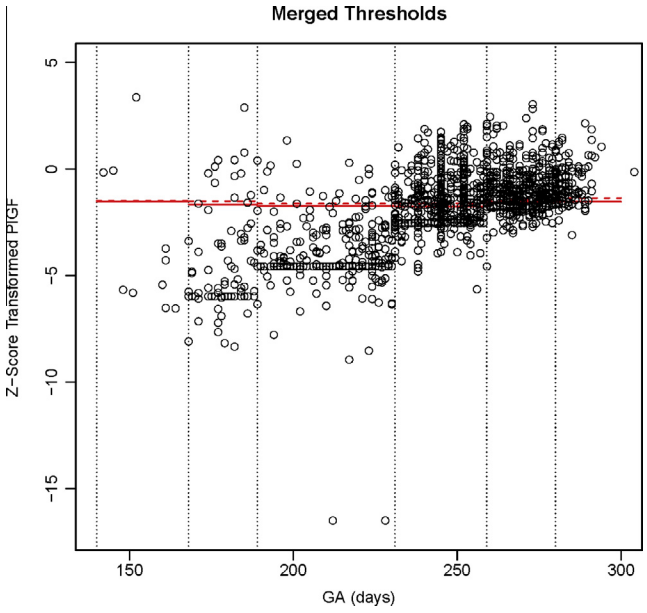

Fig. 2. PlGF Z-Scores from cases with PlGF Z-Score thresholds.

threshold. PlGF Z-Scores from cases are shown in Fig. 2 with the merged thresholds based on the non-case PlGF Z-Scores (PlGF MoMs plot similar, but not shown). The highest rates of case identification occurred prior to 33 weeks' gestation, with loss of sensitivity in sampling at later GA (Table 4) which is consistent with published reports [3,6]. (Note that the four platforms are anonymized by randomly assigned platform numbers 1–4 in Table 4). From the figures reported in Table 4, it is clear that the merged thresholds (based on Z-Score or MoM transformation) performed similarly. The Z-Score merged thresholds yielded rates of incorrect diagnosis averaging 5% (range 2.6–8.3%) as expected by construction of reference curve (note that the lower bound of the 95% confidence interval for the merged thresholds resulted in an average of 6.2% false positives with range 5–12%). The MoM merged threshold performance showed an average of 10% non-cases incorrectly identified as cases (range 3.2–32%).

4. Discussion

We have developed and validated a generalizable method for pooling and merging laboratory data from four different analytical platforms. We illustrate our pooled data analysis strategy with hypertensive disorders of pregnancy and the circulating maternal biomarker PlGF, which is increasingly used in clinical practice.

Merging of PlGF measurements quantified on different assay platforms allowed for the development of a best reference curve (BRC) for uncomplicated pregnancy that can be applied, in future, to the identification of complicated pregnancies. We would like to highlight three main components of the above analyses. Firstly, using clustering analysis and the Adjusted Rand Indices, we compared the ability of each transformation (Z-Score and MoM) to produce measurements that are easily merged. Both transformations performed well. The rates of case identification using reference thresholds derived from the merged data (Table 4) indicated that the MoM transformation performed slightly better than the Z-Score transformation in later age groups. The choice between using the MoM BRC-PlGF or the Z-score BRC-PlGF may rely on the investigator's preferences. The differences between the overall diagnostic rates of both merged thresholds and the average diagnostic rates of the platform-specific thresholds were not statistically significant at the 5% level for any GA category. We conclude that the merged BRC performs just as well as those estimated specifically for each platform, with the added practical advantages of being based on a much larger and broader sample. The merged BRC is particularly useful for collaborative investigations across cohorts.

Our method has been developed to study pregnancy-related screening or diagnosis data but is applicable to any medical condition where there is intrinsic variability in the tests that measure the same biomarker(s). The choice of BRC for any given study and any biomarker will in general depend on the distribution of the data being used. It is clear from our results in this PlGF merging study of pregnancy blood samples that the diagnostic information itself has not been degraded by merging the data from these heterogeneous platforms and cohorts.

Of the possible biases in our study, some are intrinsic to constructing reference ranges whether from single or multiple datasets. They are considered no further than to say that the validity of our approach depends on our definition of non-cases and the requirement that no non-case contributed more than one value to the BRC. Unstandardized use of the same analytical platform in different laboratories may lead to small systematic biases in the results. We found no evidence for this kind of bias of a magnitude that could constitute an important problem in our application here, as removal of any single cohort in our LOO-CV methodology (and the respective PlGF measurements) did not significantly alter the merged BRC-PlGF across all GA ranges. Distributions of both serum and plasma measurements were shown to be similar and therefore these data were combined for the analysis, however the combination of matrices may be a potential source of variability.

Table 4
Rates of correct identification of cases for merged and platform-specific (P-S) thresholds [and lower bound of associated 95% CI].

GA category		$20–23^{+6}$		$24–26^{+6}$		$27–32^{+6}$		$33–36^{+6}$		$37–39^{+6}$		40+	
Method	Platform												
Z-Score	1	1.00	[1.00]	1.00	[1.00]	0.95	[0.97]	0.16	[0.20]	0.06	[0.10]	0.00	[0.00]
	2	1.00	[1.00]	1.00	[1.00]	0.96	[0.97]	0.73	[0.79]	0.38	[0.43]	0.25	[0.28]
	3	0.71	[0.71]	0.71	[0.79]	0.81	[0.81]	0.46	[0.49]	0.46	[0.49]	0.08	[0.13]
	4	0.00	[0.00]	0.45	[0.55]	0.92	[0.92]	0.27	[0.31]	0.37	[0.40]	0.20	[0.33]
	Overall	0.67	[0.67]	0.83	[0.86]	0.91	[0.92]	0.44	[0.48]	0.34	[0.38]	0.17	[0.23]
MoM	1	1.00	[1.00]	1.00	[1.00]	0.94	[0.97]	0.13	[0.15]	0.00	[0.00]	0.00	[0.00]
	2	1.00	[1.00]	1.00	[1.00]	0.97	[0.97]	0.84	[0.86]	0.59	[0.64]	0.28	[0.31]
	3	0.71	[0.71]	0.71	[0.79]	0.79	[0.81]	0.29	[0.37]	0.21	[0.34]	0.00	[0.04]
	4	0.00	[0.00]	0.55	[0.55]	0.92	[0.92]	0.25	[0.31]	0.35	[0.42]	0.31	[0.37]
	Overall	0.67	[0.67]	0.84	[0.86]	0.91	[0.92]	0.41	[0.46]	0.35	[0.41]	0.20	[0.24]
P-S thresholds average rate		0.68	[0.68]	0.81	[0.83]	0.91	[0.92]	0.42	[0.53]	0.34	[0.41]	0.20	[0.32]

It is assumed that there is statistical independence between platforms and case-mix. In an ideal situation, each blood sample from the pregnant woman would be measured on all four platforms and in the same laboratory to allow direct comparison. However in this application, to our knowledge, PlGF measurements of this form are currently not available for any pregnancy cohorts.

Biases that could be generated during the merging of the PlGF data are highly relevant because we wanted to establish a generalizable method applicable to merged data of many tests in multiple contexts. After merging the non-case values, there was no residual cluster that could be attributed to one analytical platform. By LOO-CV we also excluded the possibility of a dominant contribution from any single cohort. In relation to this specific study, since we did not systematically seek every known pregnancy cohort globally and also since relevant pregnancy cohorts with angiogenic factor analyses from low and middle income countries are lacking, there is a possibility of potential inclusion bias. We believe any such bias to be minimal but this will be addressed more specifically in our future clinical analyses of this database.

The accuracy of medical diagnostic tests is usually considered [7] in relation to a clear, verifiable, single diagnosis using laboratory methods that are already standardized for the purpose of introduction into routine clinical practice. Examples include tests in prenatal screening for fetal chromosomal abnormalities [8]. The present paper describes a methodology which may be more applicable to research and discovery. Preeclampsia, like other syndromes such as many inflammatory and cardiovascular conditions, lacks sharp diagnostic definitions. In preeclampsia and related placentally-mediated disorders of pregnancy, there are many clinical presentations in a gray zone between normality and abnormality. Furthermore the disease can be extremely variable, indicating underlying heterogeneity. Clinical diagnosis could be improved in relation to better definitions of disease subtypes using biomarkers (3), but further discovery relating to diagnostic challenges is impeded by the low power of single studies, and the fact that researchers have used several laboratory assays with differing analytical performances for measuring single biomarkers. It is at this level that our method is likely to be most useful.

We have constructed a "PlGF converter" that enables any researcher or clinician to calculate a general PlGF percentile based on a PlGF concentration measured during pregnancy after week 20, on any of the four platforms included in our study (It also includes options for comparison against merged thresholds (both Z-Score and MoM) and platform-specific thresholds). A link to the PlGF converter is on the CoLAB home page (http://pre-empt.cfri.ca/co-laboratory/global-pregnancy-collaboration).

We hope that the awareness of our merging method will encourage researchers to plan their studies, in any biomarker field, to allow their data to be merged with other future datasets in order to gain more statistical power and research value for rare study outcomes. Harmonization of data collection is being addressed now in various clinical arenas [9], also for preeclampsia [10,11], where application of the new "PlGF converter" could be useful when merging pregnancy PlGF biomarker studies.

The clinical value of our PlGF reference curve has yet to be validated. Future work will use the merged BRC-PlGF across gestational age to better characterize clinical subgroups of the hypertensive disorders of pregnancy and other important complications such as fetal growth restriction. We will use the BRC-PlGF to explore how different clinical groups of HDP and fetal growth restriction are sub-classified on the basis of PlGF [3]. We will extend our studies, using similar merging algorithms, to other angiogenic markers (sFlt-1 and sEng), and to other pregnancy conditions such as gestational diabetes mellitus and more rare pregnancy outcomes, including intrauterine fetal death. Pregnancy disorders are not the only relevant conditions where this strategy could be useful: other complex syndromes such as the metabolic syndrome or polycystic ovarian syndrome present the same problems.

Overall, we show here that heterogeneity of laboratory assays in biomarker studies need not be a barrier for inclusion in pooled analysis of individual patient datasets. The method shows how to transform and validate merging for any set of biomarkers obtained from different laboratory platforms in any field.

Acknowledgements

We would like to acknowledge the contributions of Drs. Zhang (member of the Global CoLaboratory group) and Schisterman who provided the NIH study data. Dr. Zhang and Schisterman were supported by the Intramural Research Program of the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, National Institutes of Health.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.preghy.2015.12.002.

References

[1] Task Force on Hypertension in Pregnancy, Hypertension in pregnancy, Report of the American College of Obstetricians and Gynecologists' Task Force on Hypertension in Pregnancy, Obstet. Gynecol. 122 (2013) 22–31.
[2] R.B. Ness, J.M. Roberts, Heterogeneous causes constituting the single syndrome of preeclampsia: a hypothesis and its implications, Am. J. Obstet. Gynecol. 175 (1996) 1365–1370.
[3] A.C. Staff, S.J. Benton, P. von Dadelszen, J.M. Roberts, N. Taylor, R.W. Powers, et al., Redefining preeclampsia using placenta-derived biomarkers, Hypertension 61 (2013) 932–942.
[4] R.N. Taylor, J. Grimwood, R.S. Taylor, M.T. McMaster, S.J. Fisher, R.A. North, Longitudinal serum concentrations of placental growth factor: evidence for abnormal placental angiogenesis in pathologic pregnancies, Am. J. Obstet. Gynecol. 188 (2003) 177–182.
[5] S.J. Benton, Y. Hu, X. Fang, K. Kupfer, S.W. Lee, L.A. Magee, P. von Dadelszen, Can placental growth factor identify placental intrauterine growth restriction in small for gestational age fetuses?, Am J. Obstet. Gynecol. 206 (2012) 1–7.
[6] R.J. Levine, C. Lam, C. Qian, K.F. Yu, S.E. Maynard, B.P. Sachs, et al., Soluble endoglin and other circulating antiangiogenic factors in preeclampsia, N. Engl. J. Med. 55 (2006) 992–1005.
[7] J.B. Reitsma, K.G. Moons, P.M. Bossuyt, K. Linnet, Systematic reviews of studies quantifying the accuracy of diagnostic tests and markers, Clin. Chem. 58 (2012) 1534–1545.
[8] S.K. Alldred, J.J. Deeks, B. Guo, J.P. Neilson, Z. Alfirevic, Second trimester serum tests for Down's syndrome screening, Cochrane Database Syst. Rev. (2012) 13–16.
[9] D. Doiron, P. Burton, Y. Marcon, A. Gaye, B.H. Wolffenbuttel, M. Perola, et al., Data harmonization and federated analysis of population-based studies: the BioSHaRE project, Emerg. Themes Epidemiol. 10 (2013) 12.
[10] L. Myatt, C.W. Redman, A.C. Staff, S. Hansson, M.L. Wilson, H. Laivuori, et al., Strategy for standardization of preeclampsia research study design, Hypertension 63 (2014) 1293–1301.
[11] G.J. Burton, N.J. Sebire, L. Myatt, D. Tannetta, Y.L. Wang, Y. Sadovsky, et al., Optimising sample collection for placental research, Placenta 35 (2014) 9–22.