Improving the quality of anthropometric measures during medical consultations with children aged under five years old in Burkina Faso

F

5	
6	Aziza Merzouki ^{a,*,#} , Wessel Valkenburg ^{b,#} , Marc Bayala ^c , Maroussia Roelens ^a , Olivia Keiser ^a ,
7	Amara Amara ^b
8	^a Institute of Global Health, University of Geneva, Geneva, Switzerland
9	^b Terre des Hommes Foundation, Lausanne, Switzerland
10	^c Terre des Hommes Foundation, Ouagadougou, Burkina Faso
11	
12	* <u>Corresponding author:</u>
13	Aziza Merzouki, PhD
14	Institute of Global Health, University of Geneva
15	Chemin des Mines 9, 1202 Geneva, Switzerland
16	Tel. +41 78 712 56 46
17	FatmaAziza.Merzouki@unige.ch
18	
19	* <u>Alternate corresponding author:</u>
20	Amara Amara, PhD
21	Avenue de Montchoisi 17, 1006 Lausanne, Switzerland
22	Tel. +33 6 45 01 13 49
23	amara.amara@tdh.ch
24	
25	# these authors contributed equally
26	Word count: Abstract 242 words; main text 2994 words; 6 figures; 1 supplementary material
27	file

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

28 Abstract

Objective: Millions of medical consultations are conducted each year in Burkina Faso using the Electronic Register of Consultations (REC). Based on the consultation data collected, we present a method to quantify the quality of individual and ensembles of consultations conducted by frontline healthcare workers (FHWs).

33

Methods: We focus on anthropometric measurements and vital signs (age, weight, height, mid-upper arm circumference and temperature) of children aged between two months and five years old. We compare individual and ensemble of consultations to a multivariate probability distribution defined by an external population-specific, gold standard consultation dataset. By comparing the distributions of consultations to the reference probability distribution, we define a score to rate the quality of measurements and data entry of each FHW.

41

Findings: The defined scores allow us to detect which measurements are most problematic. They also allow us to detect potential biases in the consultation and treatment of different patient groups. No systematic gender-bias was found among FHWs. Height measurements were the most challenging; consultations with the lowest scores were associated with underestimated heights in children. Among these consultations, height was found to be even more underestimated among boys than girls.

48

49 Conclusion: Our findings enable us to support capacity building of frontline healthcare 50 workers. The REC can be enriched with real-time specific alert on errors, individual FHW can 51 be proposed targeted trainings, and dynamic dashboards can support district managers to 52 navigate the entire population of FHWs and understand which problems should be prioritised. 53

54 Research in context

55 Knowledge before this study

56 The use of the Electronic Register of Consultations (REC) improved Frontline Healthcare 57 Workers' (FHWs) adherence to the Integrated Management of Childhood Illness (IMCI) 58 guidelines at the primary care level in Burkina Faso. The improvement included a better 59 identification of danger signs and an increase in the proportion of correctly classified children under five years old. A former study reported how FHWs perceived the use and impact of the 60 61 REC in their daily practice. While a high degree of satisfaction was expressed, FHWs also 62 proposed improvements. FHWs proposed to increase the frequency of supervision and 63 evaluation visits, which usually take place every three months. Supervision from district teams and coaches was globally positively perceived by FHWs, as it allowed them to identify and 64 address errors, and therefore helped them to learn and improve. FHWs also proposed 65 66 receiving compensations or prizes for the best health centres according to the evaluations.

67 Contribution of this study

68 In this study, we proposed a method to assess the quality of consultations conducted by 69 FHWs. We focused on anthropometric measurements and vital signs that are systematically 70 measured by FHWs during consultations of children aged between two months and five years 71 old. We showed how this method can feed a live alert system that invites FHWs to verify their 72 input in-real time when potential errors in specific measurements or data entries are 73 identified. We found that height (length) measurements of children were the most 74 challenging, as height (length) was frequently underestimated. Finally, we presented a dynamic dashboard that informs health district managers on the quality of care across the 75 76 country (using a medal reward system), so they can prioritize their interventions and provide FHWs with targeted support to improve their skills. 77

78 Introduction

The burden of under-five mortality in West Africa is high. Out of 1000 live births in Burkina Faso in 2019, 87 children died before the age of five (1); the leading causes of these deaths are preventable or treatable conditions (2,3). In comparison, the under-five mortality rate in Europe and Northern America was five deaths per 1000 live births in 2019 (4). To reduce the morbidity and mortality of children aged under five years in low- and middle-income countries, WHO and UNICEF developed the Integrated Management of Childhood Illness (IMCI) protocol, which supports the combined treatment of major childhood illnesses (5,6).

86

87 The paper-based IMCI was digitalized by Terre des hommes (Tdh) and the Ministry of Health (MoH) in Burkina Faso, who co-created leDA (the Integrated e-Diagnostic Approach) to 88 89 improve adherence to IMCI. IeDA includes the Electronic Register of Consultations (REC), a mobile application that runs on Android-based tablets and guides frontline healthcare 90 91 workers (FHWs) through IMCI to diagnose sick children. The deployment of IeDA in Burkina 92 Faso has been growing steadily, covering 67% of all primary healthcare centres (PHCs) in the 93 country by 2021. It is used by thousands of FHWs during the consultations of millions of 94 children each year.

95

Using machine learning, important work was done to leverage the large amount of
consultation data collected with IeDA (7). IeDA data were analysed to: (i) predict epidemic
outbreaks(8,9), (ii) deploy smart dashboards to inform and support decision makers, and (iii)
build FHWs capacity to improve the quality of care (10).

100

During a consultation using the REC, the FHW answers a series of questions related to the child's anthropometry, vital signs, clinical signs and reported symptoms. The accuracy of the FHW's answers and his input data are key to a high-quality consultation; the consultation process, the final diagnostic classification and the recommended treatment depend on these data. Very few FHWs who conduct consultations are medical doctors; most of them are nurses, itinerant healthcare workers and midwives. FHWs using the REC are keen to get feedback on their performance and willing to know how they compare to their peers (11).

108 Moreover, managers of PHCs and health districts show a strong interest to be continually 109 informed on how FHWs perform and how they could support their work.

110

To define what is a high- or low-quality consultation, one needs to define a reference. We can use the REC database itself as a reference and compare each FHW's input to the typical input over the entire country. This would be equivalent to an anomaly (outlier) detection. However, a weakness of this approach is the possibility of systematic errors which are made by all FHWs collectively and may lead to potential biases. The other possibility, which we use in this study, is to consider an external database consisting of consultations made by experts, which may serve as a reference (12,13).

118

119 We focus on anthropometric measurements, which are essential to assess the growth and 120 nutritional status of the child (3,14), and on vital signs (temperature), that are systematically 121 measured by FHWs during consultations of children aged between two months and five years 122 old. We propose a method to quantify the quality of a single consultation and an ensemble of 123 consultations conducted by FHWs. We use this method to feed a live alert system that 124 identifies potential errors in specific measurements or data entries. We explore if this method 125 reveals any bias in the treatment of boys versus girls. Finally, we present a dynamic dashboard that provides managers with an overview on the quality of care across the 126 127 country so they can support FHW improve their skills.

128 Methods

129 Data

Data from IeDA includes over nine million consultations, increasing by approximately 200'000 consultations per month. During each consultation, the following information is systematically recorded in real time: age, height (length in case of infants), weight, mid-upper arm circumference (MUAC) and temperature. Z-scores of weight-for-height, height-for-age and weight-for-age are calculated and also recorded. These z-scores are based on the WHO worldwide standard model for growth of children(15). Our main analysis includes data from all consultations conducted during 2020 using the REC, across the country.

137

To rate the quality of consultations, we use an external database as a reference. We use 138 139 aggregated statistics from data collected during an earlier project (12,13,16), which we will 140 refer to as `the audit data' from here on. This audit data was collected between 2014-2017 by 141 specifically trained IMCI experts during the consultations of children aged between two 142 months and five years old, in eight districts (Solenzo, Nouna, Dedougou, Boromo, Toma, 143 Goursi, Ouahigouya and Titao) in Burkina Faso. The aim of this former study was to determine 144 whether IeDA increases adherence to IMCI and hence improves the quality of care for children 145 in PHCs. Due to data-protection, we could not work with the raw audit data. We could 146 however access the means of weight-for-height, height-for-age and weight-for-age z-scores, 147 temperature and MUAC, as well as the covariance matrix of these five variables.

148 Reference model

We make the hypothesis that the distribution of patients over the five-dimensional real space
of weight-for-height, height-for-age and weight-for-age z-scores, temperature and MUAC is a
multivariate Gaussian distribution,

152

$$P(\vec{x}) = (2\pi)^{-\frac{k}{2}} |\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2}(\vec{x}-\vec{\mu})^{T}\Sigma^{-1}(\vec{x}-\vec{\mu})}$$
(1)

where \vec{x} is the *k*-dimensional vector describing a single observation, k = 5 is the number of degrees of freedom, Σ is the $k \times k$ positive-definite covariance matrix of the model (estimated from the data), $|\Sigma|$ is the determinant of the matrix Σ , and $\vec{\mu}$ is the vector 156 containing the mean values of the k variables. The reference model of 'good quality of care' 157 is then fully defined once $\vec{\mu}$ and Σ are known.

158 Individual consultations

We first assign a score to individual consultations (data points in the space of anthropometric 159 160 z-scores, temperature and MUAC); representing the probability that the combination of 161 entered values is accurate. This score allows us to raise a live alert to the FHW when erroneous input data is suspected. With the multivariate Gaussian reference model, the probability 162 density associated with an individual consultation, is given by Equation 1. We trigger an alert 163 when the probability density is lower than a threshold, corresponding to the worst 10% 164 consultations in the database for year 2020. The threshold was set to avoid doing harm to the 165 166 willingness of FHWs to use the tool at all.

167

168 The second step consists in identifying the most suspicious input data in order to guide the 169 FHW in real time to check specific measurements and entries. Therefore, we analyse 170 separately the five variables MUAC, temperature and three z-scores, denoted as x_i , with i =171 1 ... 5, and compute individual variable scores $score_i = \frac{x_i - \mu_i}{\sigma_i}$, given the reference experts' 172 means μ_i and standard deviations σ_i . Any x_i is suspect when $|score_i| > thresh_i$, with 173 $thresh_i = 3, \forall i$. When $score_i < -thresh_i$ or $score_i > thresh_i$, x_i is considered extremely 174 low or extremely high, respectively.

175 When MUAC and temperature are identified as extreme ($|score_{MUAC}| > thresh_{MUAC}$ and 176 $|score_{temperature}| > thresh_{temperature}$), the FHW is directly informed by an alert message 177 in the REC and invited to check the input values. Weight, height and age are identified as being 178 abnormally low or high based on the combination of scores related to weight-for-age, height-179 for-age and weight-for-height (Supplementary Materials).

180 Ensembles of consultations over a given time period

181 Probability

182 We projected consultations \vec{x} on a new five-dimensional orthogonal space (with

uncorrelated dimensions); $\vec{x} \rightarrow \vec{y} = U^{-1}(\vec{x} - \vec{\mu})$, where the column vectors of U are the

- 184 (normalized) eigenvectors of the reference covariance matrix Σ .
- 185 For an ensemble of *N* consultations (done by a single FHW, in a single PHC) over a given time
- 186 period (*N* draws from the reference distribution), the measured sample mean $\hat{\mu}_{N,i}$ of a single
- 187 component y_i , and its measured sample variance $\hat{s}_{N,i}^2$ are distributed as Gaussian and χ^2_{N-1}
- 188 respectively.
- 189 The probability density of an ensemble $\{\vec{y}\}_N$ of N consultations, is then given by,

190
$$P(\{\vec{y}\}_N) = \prod_i P(\hat{\mu}_{N,i}) P(\hat{s}_{N,i}).$$
(2)

191 For more details, including equations of $P(\hat{\mu}_{N,i})$ and $P(\hat{s}_{N,i})$, see Supplementary Material.

 $P({\vec{y}}_N)$ is the probability that the observed ensemble ${\vec{y}}_N$ is generated by the reference 192 model. This corresponds to the p-value of the null hypothesis that the consultations are done 193 194 by an expert. The p-value informs us about the significance of a deviation from the reference model, but it does not necessarily inform us about the quality of the work. In fact, even for 195 196 the same deviation from the reference model, a large number of consultations will lead to a 197 much lower p-value than a small number of consultations (Figure A 1). Vice versa, different 198 deviations from the reference model can lead to the same p-value, if the number of 199 consultations differs accordingly.

200 Quality score

201 We aim at providing healthcare workers with an actionable feedback. The feedback consists 202 of two parts: (i) an overall performance score to indicate how the quality of work of a FHW 203 compares to their peers, and (ii) suggestions about which part of their work should be 204 improved with priority. As a proxy for the quality score, we use Equation 2, which defines the 205 probability density that an ensemble of consultations is drawn from the reference distribution (significance), where we fix the number N to a constant arbitrary value for all FHWs (Figure A 206 207 2). This way, only the mean and variance of the consultation measurements are included in 208 the calculation of the proxy for the quality score; ignoring the number of consultations that

- 209 gave rise to said mean and variance. To facilitate the interpretation of the score for any user,
- 210 we define the score as the percentile of a FHW's score proxy, relative to the entire population
- of FHWs. The score therefore reflects a percentage-based ranking, and indicates how well the
- 212 FHW performs compared to their peers.
- 213
- Figure 1 and Figure A 3 illustrate that the proxy score is more suitable to measure quality, thanthe actual probability of these ensembles.

216 Gender Bias

In order to identify potential gender bias in measurements or data entry, we first compared
the typical variation in the scores of individual FHWs over different time periods (quarters of
a year), to the typical variation in their scores between boys and girls in the same time period.
We also compared the distribution of scores and the frequency of suspicious input values
(extremely low/high height, weight, MUAC, etc.) between girls and boys over all FHWs.
Statistical significance was assessed using a two-proportion z-test.

223 Results

224 Detecting common problems

225 The following analysis includes 2,042,545 consultations (939,761 girls, and 1,202,784 boys) 226 done in 2020, by 7,234 FHWs, in 1,150 PHCs from 39 districts. We compared the distributions 227 of age, temperature, MUAC and the three anthropometric z-scores, i.e., height-for-age, 228 weight-for-height and weight-for-age between the most suspicious consultations, which 229 obtained the lowest scores (worst 0.3%), and the remaining (99.7%) consultations that got 230 higher scores. The distributions are presented in Figure 2. Compared to the consultations with higher scores, the lowest score consultations had lower z-scores of height-for-age 231 232 (median(IQR) -8.7(-7.5-(-9.8)) vs -2.1(-1.0-(-3.5)), higher z-scores of weight-for-height (median(IQR) 9.0(12.3-6.9) vs -0.2(0.8-(-1.1))), and were more frequent among younger 233 234 children; 71.6% of consultations with the lowest 0.3% scores involved children under 20 235 months old vs 40.8% among the remaining 99.7% of consultations. Consultations that involved 236 overweight children (weight-for-age z-score > 2) were also more frequent among the consultations with the lowest scores (15.8% vs 0.5% among consultations with higher scores). 237

238 When identifying the specific inputs (among age, weight, height, MUAC and temperature) 239 with extreme (low or high) values, height underestimation was the most frequent, followed 240 by weight overestimation. Consultations with the 0.3% least likely combinations of 241 anthropometric and vital signs entries underestimated height in 88% of cases and 242 overestimated weight in 14%; see Figure 3 (left). These numbers correspond to 29% of height underestimation and 2% of weight overestimation among the 10% consultations with the 243 244 lowest scores. Figure 3 (right) presents the evolution of the frequency of underestimated and overestimated input variables as a function of the selected score threshold (varying from the 245 246 lowest 0.3% to the lowest 40% scores). Reproducing the analysis for consultations conducted 247 in 2019 and 2018 provided similar results (Supplementary Material).

248 Gender bias in primary healthcare

Figure 4 shows that the inter-seasonal variation in quality of care is larger than the variation between boys and girls, as the former is more skewed to the right. We can therefore not demonstrate any systematic gender bias. Considering ensemble of consultation scores of individual FHWs, Figure A 4 shows that only FHWs with fewer consultations had a larger difference in scores between girls and boys. Figure A 5-7 show how the large discrepancy in scores can be explained by outliers having a large impact on the score when the sample size is limited (<1000 consultations).

Despite the similar overall distribution of scores between girls and boys, the lowest scores for boys were even more extreme (lower) than for girls (see Figure A 8). Among the consultations with the lowest scores (0.3%; 2,818/939,761 girls; 3,316/1,202,784 boys), the most significant difference between the consultations for girls and boys with the lowest scores involved height measurement (Figure 5). Height in girls was less often underestimated than height in boys (84% for girls; 90% boys; p<0.01).

262 Live alerts

The REC was expanded to estimate in real time the score of anthropometric and vital signs values that are systematically entered by a FHW during a consultation. When a consultation gets a low score, suspicious values (extremely low or high) are identified, and an alert message invites the FHW to check their entries. In case of a confirmed measurement or data entry

267 error, the FHW can correct the input value and continue the consultation. Figure A 9 presents268 screenshots of the expanded REC application.

269 Dynamic dashboards for managers

270 A dynamic dashboard was implemented (using Tableau Desktop 2020.3 software) and made 271 accessible online to provide health centre and district managers with an up-to-date overview 272 of the PHC score distributions across the country (see Figure 6). In this dashboard, each PHC 273 was represented by the ensemble of all consultations (focusing on anthropometry and temperature data entries) conducted in this site. Depending on the scores obtained, PHCs 274 275 were categorized in three categories: gold, silver and bronze (details in Supplementary 276 Material). To help identify the measurements that are the most suspicious in a selected PHC 277 and that lead to a lower quality score, the distribution of all consultations of the PHC are 278 displayed (as a scatter plot) together with the reference model (orange lines) for comparison. 279 In a separate panel, the proportion of gold, silver and bronze medals per health district are 280 also presented. Districts in the Northern and Eastern parts of Burkina Faso (e.g., Dori, Diapaga) 281 had higher percentages of bronze PHCs, while districts in the Western part of the country (e.g., 282 Tougan, Tenado) had a majority of silver and gold PHCs.

283 Discussion

Using a multivariate gaussian model, we quantified the quality of anthropometric measures and vital signs as entered by FHWs during consultations of children aged from two months to five years old. The quality measure of individual consultations was used to feed a live alert system that identifies suspect input values of age, height, weight, MUAC and temperature. The quality measure of ensembles of consultations were used to identify suspect behaviours among FHWs based on all the consultations they have conducted during a specific time period.

Among the lowest score consultations, height measurements were the most problematic, as height was frequently underestimated. The challenge when measuring a child's height (length) lies in the difficulty to keep him fully stretched and still, especially when the child is crying and struggling (19). Among the consultations with the lowest scores, height was significantly more underestimated for boys than for girls, which may be associated to higher activity levels in boys (20–22).

297

298 Bronze medals were more frequent in the Northern and Eastern districts of Burkina Faso, 299 while Western Districts had a majority of silver and gold medals. In addition to differences in 300 the quality of anthropometric and vital signs measures, this geographical heterogeneity could 301 also reflect variations in socioeconomic and nutritional status across the country (17,18). 302 Previous reports showed that food insecurity is the most prevalent in the Sahel (Northern) 303 and East regions. If these geographical differences and local child population characteristics 304 are confirmed, the reference model will be adapted accordingly for a more accurate 305 estimation of FHWs' quality of work.

306

307 Our analyses have some limitations. As a reference of what is a high-quality consultation, we 308 used aggregated statistics of experts consultation data (12,13,16). This choice has 309 weaknesses, including that (i) the statistics were computed over a finite number of 310 consultations (order of 1500), and that the reference distribution is only measured up to a few 311 percent accuracy, (ii) the data were collected in a different period than the period in which we 312 tested the quality of care, (iii) the data covered a subset of districts only, (iv) the data were 313 collected by only two trained nurses, which may lead to biased results due to the limited 314 sample size.

315

316 In collaboration with the MoH, Tdh has initiated the regular collection of consultations 317 performed by recognized experts, in order to recalibrate the reference model and keep it as 318 close to the real population as possible. Data collection will be done over a heterogeneous 319 sample of the population, and cover different regions and seasons. As we worked with 320 aggregated audit data, we made the simplifying hypothesis that the distribution of patients 321 over the five-dimensional real space of anthropometric and vital signs variables follows a 322 multivariate gaussian distribution. The validity of this assumption and of the audit data will be tested in the future, when the above-mentioned dataset of expert consultations becomes 323 324 available. As a future work, we will further explore potential factors associated with lower 325 consultation scores (e.g., FHW's workload, time at which data are entered in the REC), and 326 focus on the quality of reported symptoms and classifications.

327 Conclusion

By developing a method that uses a large dataset to qualify the healthcare provided to children in a low-resource setting, we showed how we can foster capacity building through automated personal live feedbacks and dynamic dashboards informing district managers' decision-making, and explore potential biases in the quality of healthcare.

332 Acknowledgements

- 333 The authors thank Antoine Geissbuhler, Seydou Toguiyeni, Noël Nacoulma, Noël Zonon,
- 334 Florian Triclin, Iveth Gonzalez, Sandrine Busiere, Riccardo Lampariello, the IeDA team of
- 335 Terre des hommes in Burkina Faso and the Ministry of Health from Burkina Faso for their
- input and fruitful discussions.

337 Funding

- 338 This work was funded by the Cloudera Foundation, and technically supported by the
- 339 Cloudera Foundation and the Tableau Foundation.

340 Conflict of interest

341 We declare no competing interests.

342 References

- 343 1. Burkina Faso (BFA) Demographics, Health & Infant Mortality [Internet]. UNICEF
- 344 DATA. [cited 2021 Mar 3]. Available from: https://data.unicef.org/country/bfa/
- Sanyang Y. Prevalence of under-five years of age mortality by infectious diseases in
 West African region. Int J Afr Nurs Sci. 2019 Jan 1;11:100175.
- Caulfield LE, de Onis M, Blössner M, Black RE. Undernutrition as an underlying cause
 of child deaths associated with diarrhea, pneumonia, malaria, and measles. Am J Clin Nutr.
 2004 Jul;80(1):193–8.
- 4. UNICEF, WHO, World Bank Group. Levels and Trends in Child Mortality [Internet].
- 351 2020 [cited 2021 Apr 13]. Available from: https://www.unicef.org/media/79371/file/UN-
- 352 IGME-child-mortality-report-2020.pdf.pdf
- Gera T, Shah D, Garner P, Richardson M, Sachdev HS. Integrated management of
 childhood illness (IMCI) strategy for children under five. Cochrane Database Syst Rev

355 [Internet]. 2016 [cited 2021 Mar 3];(6). Available from:

356 https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD010123.pub2/full?cookies

357 Enabled

358 6. Gove S. Integrated management of childhood illness by outpatient health workers:

359 technical basis and overview. The WHO Working Group on Guidelines for Integrated

360 Management of the Sick Child. Bull World Health Organ. 1997;75 Suppl 1:7–24.

361 7. Large-scale digitalized IMCI in Subsaharan Africa: on how AI can help decrease child

362 mortality & morbidity, and predict epidemic outbreaks. [Internet]. [cited 2021 Feb 15].

363 Available from: https://gdhf2019.dryfta.com/program-schedule/program/65/artificial-

364 intelligence-and-machine-learning-applications-in-global-health

3658.Faster reaction to epidemics: toward predicting outbreaks in Burkina Faso - AMLD

366 [Internet]. [cited 2021 Mar 2]. Available from: https://appliedmldays.org/highlights/84

367 9. Harvey D, Valkenberg W, Amara A. Predicting Malaria Epidemics in Burkina Faso With

368 Gaussian Processes [Internet]. Rochester, NY: Social Science Research Network; 2021 Feb

369 [cited 2021 Mar 29]. Report No.: ID 3786697. Available from:

370 https://papers.ssrn.com/abstract=3786697

10. Improving the quality of child consultations in Burkina Faso using ML: detecting and

addressing Health Workers' diagnostic mistakes - AMLD [Internet]. [cited 2021 Mar 2].

373 Available from: https://appliedmldays.org/highlights/85

11. Bessat C, Zonon NA, D'Acremont V. Large-scale implementation of electronic

375 Integrated Management of Childhood Illness (eIMCI) at the primary care level in Burkina

Faso: a qualitative study on health worker perception of its medical content, usability and

impact on antibiotic prescription and resistance. BMC Public Health. 2019 Apr 29;19(1):449.

378 12. Blanchet K, Lewis JJ, Pozo-Martin F, Satouro A, Somda S, Ilboudo P, et al. A mixed

379 methods protocol to evaluate the effect and cost-effectiveness of an Integrated electronic

380 Diagnosis Approach (IeDA) for the management of childhood illnesses at primary health

facilities in Burkina Faso. Implement Sci IS [Internet]. 2016 Aug 4 [cited 2021 Feb 11];11.

382 Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4973038/

13. LSHTM. IeDA for the management of illness in under-five children at the primary

health care level in Burkina Faso: Findings from a stepped-wedge cluster randomised trial

385 [Internet]. 2018 [cited 2021 Feb 11]. Available from:

386 https://www.tdh.ch/sites/default/files/ieda_brochure.pdf

- 387 14. WHO. Nutrition Landscape Information System (NLIS) country profile indicators:
- 388 interpretation guide [Internet]. 2010 [cited 2021 Mar 8]. Available from:
- 389 https://www.who.int/nutrition/nlis_interpretation_guide.pdf
- 390 15. The WHO Child Growth Standards [Internet]. [cited 2021 Apr 13]. Available from:
- 391 https://www.who.int/tools/child-growth-standards/standards
- 392 16. Sarrassat S, Lewis JJ, Some AS, Somda S, Cousens S, Blanchet K. An Integrated
- 393 eDiagnosis Approach (IeDA) versus standard IMCI for assessing and managing childhood
- 394 illness in Burkina Faso: a stepped-wedge cluster randomised trial. BMC Health Serv Res. 2021

395 Apr 16;21(1):354.

- 396 17. World Bank Group. Burkina Faso Poverty, Vulnerability, and Income Source
- 397 [Internet]. 2016 Jun [cited 2021 Mar 24]. Report No.: 115122. Available from:
- 398 http://documents1.worldbank.org/curated/en/392811495031260225/pdf/Burkina-Faso-
- 399 poverty-and-vulnerability-analysis.pdf
- 400 18. Zon H, Pavlova M, Groot W. Regional health disparities in Burkina Faso during the
- 401 period of health care decentralization. Results of a macro-level analysis. Int J Health Plann
- 402 Manage. 2020;35(4):939–59.
- 403 19. WHO MULTICENTRE GROWTH REFERENCE STUDY GROUP, Onis M. Reliability of
- 404 anthropometric measurements in the WHO Multicentre Growth Reference Study: WHO
- 405 Multicentre Growth Reference Study Group. Acta Paediatr. 2007 Jan 2;95:38–46.
- 406 20. Ellis L, He P. Sex differences in fetal activity and childhood hyperactivity. Res J Dev
 407 Biol. 2014 Apr 2;1(1):1.
- 21. Campbell DW, Eaton WO. Sex differences in the activity level of infants1. Infant Child
 409 Dev. 1999;8(1):1–17.
- 410 22. Almli CR, Ball RH, Wheeler ME. Human fetal and neonatal movement patterns:
- 411 Gender differences and fetal-to-neonatal continuity. Dev Psychobiol. 2001;38(4):252–73.
- 412



414

415 Figure 1 2D projection of the consultations of four Frontline Healthcare Workers (FHWs), over Quarter 4 of 2020. Each dot 416 represents one patient. The yellow reference contour is the 99% confidence level contour of the reference model which was 417 derived from the audit data. Good quality healthcare is reflected by ensembles of consultations which are scattered within the 418 reference contour. In these selected examples, the FHWs on the left clearly make repeated errors in the weight and height 419 measurements of the patients. Our definition of score reflects correctly the difference in quality of care: ranked by score proxy, 420 the FHWs on the left scores the worst (3% on bottom-left; and 36% on top-left), while the FHWs on the right scores best (67% 421 on bottom-right; and 93% on top-right). However, due to the lower number of consultations (n=106) for the FHW in the lower 422 left panel, the probability of its ensemble is the same as for the higher-scoring FHW in the lower right panel (with 2167 423 consultations): ranked by probability both FHW's are in the worst 17%. Hence, the true probability is not a useful proxy for the 424 quality score.



425

426 Figure 2 Distribution of age, temperature, MUAC and three anthropometric z-scores among the consultations with the

427 Iowest 0.3% scores (denoted as Outlier = TRUE; coloured in blue), and among the remaining 99.7% consultations with higher

428 scores (denoted as Outlier = FALSE; coloured in red).



429 Figure 3 (left) Percentage of extreme anthropometric and vital sign input values among the 0.3% consultations with the lowest

430 scores conducted in 2020 (6,113/2,042,545 consultations). Blue and red bars represent the percentage of extremely low and

431 extremely high values, respectively. In comparison, we show in (right) the evolution of these percentages among the

432 consultations with the lowest 0.3% to 40% scores.

434



435

Figure 4 Distribution of absolute difference in scores for the same FHW in different seasons (blue) and for the same FHW in one season but separating the score for boys and girls (yellow). The inter-gender score variation is more skewed toward the

438 right.



440 Figure 5 Percentage of extreme anthropometric and vital sign values among (left panel) girls' consultations (2,818/939,761

441 consultations), and (right panel) boys' consultations (3,316/1,202,784 consultations) with the 0.3% lowest scores conducted

442 in 2020. Blue and red bars represent the percentage of extremely low and extremely high values, respectively.

443



445 Figure 6 Dynamic dashboard presenting an overview of the PHC scores distributions. Each PHC is represented by the ensemble

⁴⁴⁶ of consultations that were conducted there during Quarter 4 of year 2020. PHCs are categorized in three categories: gold,

⁴⁴⁷ silver and bronze based on the obtained score.