10.1148/ryai.2020190198

Appendix E1

Data Preprocessing and Augmentation

Images are first cropped such that all phalanges and the distal ulna and radius were within view. For two of four models (patch-based models), images are then resized to a height of 560 pixels, while preserving the aspect ratio. During training, a square patch of 224 × 224 pixels is cropped from the image, which is used as model input; the label assigned to this patch is equal to the label of the whole image. For the remaining two models (whole image–based models), images are padded and resized to 512 × 400 pixels, while preserving the aspect ratio, and the whole image is used as model input. Labels are scaled by dividing the skeletal age into months (ie, 228 months in 19 years). Image pixels are normalized from [0, 255] to [0, 1], transformed into RGB images by replicating the grayscale values 3 times, and preprocessed by subtraction of the normalized ImageNet mean and standard deviation. Standard data augmentation is employed with horizontal/vertical flips, rotations, shifts, scaling, and contrast manipulations.

Model Architecture: DenseNet121, Pretrained ImageNet Weights

Training.—

We used the AdamW optimizer with an initial learning rate of 1.0e-4 with weight decay 5.0e-4. We used a hybrid loss function equal to the mean of the mean absolute (L1) and mean squared errors (L2). The balanced hybrid loss function was constructed by creating 12-month age strata. Inverse frequency weights were assigned to each age strata. During loss computation, each individual sample loss was weighted by the sample's associated age stratum weight. Thus, the balanced hybrid loss is a weighted average of the individual sample losses in each minibatch. The learning rate was reduced by half if the validation mean absolute error (MAE) did not improve after 4 epochs. We trained for a maximum of 100 epochs with a batch size of 64 for patch-based models and 16 for whole image–based models; training was stopped if the validation MAE did not improve after 20 epochs.

Inference.—

For the patch-based models described above, 49 patches over an equally spaced 7 × 7 grid were extracted from the cropped image. Each patch was input into the trained model, and the Xth percentile for the 49 patches was used as the prediction for the whole image, where X (typically around 50) was tuned on a validation set. Final predictions were taken as the unweighted average across the two patch-based and two whole image–based models.

Appendix E2

Deming Regression

Deming regression evaluates concordance between two measures. Evidence for good concordance is achieved when the confidence limits of the intercept encompass 0, and the confidence limits of the slope encompass 1, with both indicating a lack of systematic bias. Graphically represented, perfect concordance is represented by a 45-degree line (the black line) and the observed concordance is represented by the Deming regression slope (red line); the more the red line aligns with the black line, the more concordant (Fig 2[ID]FIG2[/ID]).

Bland-Altman Analysis

Bland-Altman analysis compares two measures assessing if there are 1) systematic constant bias (red dashed line: mean difference showing one measure is systematically higher than the other) and 2) systematic trend bias (red solid line: slope showing bias exists with increasing values). A black line at 0 is the reference representing no bias (mean or slope) existed. Observations outside 95% confidence limits (blue lines) denote observations that fall out of statistical control (Fig 3[ID]FIG3[/ID]).

Table E1. Intraclass Correlation Coefficients (ICC).

	GPDL-BAAM (BL)	GPDL-BAAM	GPDL-BAAM (BS)	TDL-BAAM (BL)	TDL-BAAM	TDL-BAAM (BS)	RAD1	RAD2
CA	0.95757	0.95910	0.95943	0.96852	0.96942	0.96833	0.94554	0.93977
GPDL-BAAM (BL)		0.99883	0.99936	0.98911	0.98826	0.98939	0.98447	0.97611
GPDL-BAAM			0.99897	0.98840	0.98816	0.98952	0.98147	0.97281
GPDL-BAAM (BS)				0.98922	0.98855	0.98956	0.98382	0.97499
TDL-BAAM (BL)					0.99898	0.99877	0.97698	0.96676
TDL-BAAM						0.99895	0.97577	0.96510
TDL-BAAM (BS)							0.97588	0.96483
RAD1								0.98488

ICC (3) values between chronological age (CA), trauma hand radiograph trained deep learning algorithm (TDL-BAAM), Greulich and Pyle deep learning algorithm (GPDL-BAAM), radiologist 1 (RAD1), radiologist 2 (RAD2), with additional values for balanced loss (BL) and balanced sampling (BS) experiments.

Table E2. Deming Regression Results.

		TDL-BAAM (BL)	95% CI	TDL-BAAM (BS)	95% CI	GPDL-BAAM (BL)	95% CI	GPDL-BAAM (BS)	95% CI
CA	**Intercept**	10.571	(7.492, 13.651)	12.655	(9.523, 15.787)	13.195	(9.005, 17.385)	13.671	(9.596, 17.745)
	Slope	0.972	(0.947, 0.997)	0.954	(0.929, 0.979)	0.963	(0.933, 0.993)	0.959	(0.931, 0.988)
TDL-BAAM	**Intercept**	-0.467*	(-1.199, 0.265)	1.756	(1.022, 2.49)	2.189*	(-0.671, 5.05)	2.706*	(-0.073, 5.484)
	Slope	1.006	(1.001, 1.011)	0.988	(0.983, 0.993)	0.997*	(0.98, 1.015)	0.994*	(0.977, 1.01)
GPDL-BAAM	**Intercept**	−4.663	(-7.726, -1.6)	-2.335*	(-5.193, 0.523)	−1.931	(-2.78, -1.081)	−1.404	(-2.324, -0.485)
	Slope	1.028	(1.01, 1.047)	1.01*	(0.992, 1.027)	1.019	(1.014, 1.025)	1.015	(1.009, 1.021)
RAD1	**Intercept**	3.591*	(-0.01, 7.191)	5.781	(2.275, 9.286)	6.171	(3.061, 9.282)	6.678	(3.398, 9.958)
	Slope	0.948	(0.928, 0.969)	0.931	(0.911, 0.95)	0.94	(0.922, 0.958)	0.937	(0.918, 0.955)
RAD2	**Intercept**	7.169	(2.429, 11.91)	9.331	(4.633, 14.028)	9.69	(5.279, 14.1)	10.194	(5.684, 14.703)
	Slope	0.933	(0.908, 0.958)	0.916	(0.891, 0.941)	0.926	(0.903, 0.948)	0.922	(0.899, 0.945)
TDL-BAAM (BL)	**Intercept**			2.214	(1.519, 2.909)	2.66*	(-0.075, 5.395)	3.174	(0.521, 5.827)
	Slope			0.982	(0.977, 0.987)	0.991*	(0.974, 1.008)	0.988	(0.971, 1.004)
TDL-BAAM (BS)	**Intercept**					0.399*	(-2.252, 3.05)	0.922*	(-1.655, 3.498)
	Slope					1.01*	(0.993, 1.026)	1.006*	(0.99, 1.022)
GDLP-BAAM (BL)	**Intercept**							0.518*	(-0.198, 1.234)
	Slope							0.996*	(0.992, 1.001)

*Evidence of concordance, does not deviate statistically (contained within confidence limits).

Deming regression results (slopes, intercepts, and confidence intervals) between chronological age (CA), trauma hand radiograph trained deep learning algorithm (TDL-BAAM), Greulich and Pyle deep learning algorithm (GPDL-BAAM), radiologist 1 (RAD1), radiologist 2 (RAD2), including deep learning algorithms with balanced loss (BL) and balanced sampling (BS) sampling strategies. Results in main text Table 2 are not repeated.

Table E3. Bland-Altman Results.

		TDL-BAAM (BL)	TDL-BAAM (BS)	GPDL-BAAM (BL)	GPDL-BAAM (BS)
CA	**Intercept**	−10.67	−12.87	−13.36	−13.86
	Slope	0.03	0.05	0.04	0.04
	***P* value**	.07	< .01	.04	.02
	Mean	−6.65	−6.21	−8.01	−8
	SD	12.52	12.65	14.33	13.89
TDL-BAAM	**Intercept**	0.46	−1.77	−2.19	−2.71
	Slope	−0.01	0.01	0	0.01
	***P* value**	.04	< .01	.8	.54
	Mean	−0.44	0	−1.8	−1.79
	SD	2.46	2.52	8.26	8.14

GPDL-BAAM	**Intercept**	4.57	2.32	1.91	1.39
	Slope	–0.03	–0.01	–0.02	–0.02
	***P* value**	.01	.34	< .01	< .01
	Mean	0.49	0.92	–0.88	–0.86
	SD	8.33	7.81	2.49	2.32
RAD1	**Intercept**	–3.61	–5.89	–6.3	–6.83
	Slope	0.05	0.07	0.06	0.06
	***P* value**	< .01	< .01	< .01	< .01
	Mean	4.22	4.66	2.86	2.87
	SD	11.51	11.54	9.6	9.8
RAD2	**Intercept**	–7.26	–9.54	–9.94	–10.48
	Slope	0.07	0.09	0.08	0.08
	***P* value**	< .01	< .01	< .01	< .01
	Mean	2.78	3.22	1.42	1.43
	SD	14.57	14.79	12.44	12.71
TDL-BAAM (BL)	**Intercept**		–2.23	–2.67	–3.18
	Slope		0.02	0.01	0.01
	***P* value**		< .01	.38	.21
	Mean		0.44	–1.36	–1.35
	SD		2.7	8.05	8
TDL-BAAM (BS)	**Intercept**			–0.41	–0.92
	Slope			–0.01	–0.01
	***P* value**			.33	.55
	Mean			–1.8	–1.79
	SD			7.78	7.7
GPDL-BAAM (BL)	**Intercept**				–0.52
	Slope				<0.01
	***P* value**				.15
	Mean				0.01
	SD				1.97

Bland-Altman results (slopes, intercepts, *P* values, means) between chronological age (CA), trauma hand radiograph trained deep learning algorithm (TDL-BAAM), Greulich and Pyle deep learning algorithm (GPDL-BAAM), radiologist 1 (RAD1), radiologist 2 (RAD2), including deep learning algorithms with balanced loss (BL) and balanced sampling (BS) sampling strategies. Results in main text Table 3 are not repeated.