Health Outcomes in Mid-Ages: Multistate time to event Statistical Models versus Long Short Term Memory (LSTM) Recurrent Neural Network (RNN) Models

Lakshmi K. Raut

Visiting Fellow, University of Chicago
lakshmiraut@gmail.com

December 31, 2019

Disclaimer: The views, thoughts, and opinions expressed in the paper belong solely to the author, and do not necessarily represent the views of any institution, other group or individual.

Slides presented at the ASSA 2020 Annual Meeting, January 3 - 5, 2020, San Diego, California, USA.

Overview

Objective

- What kind of aging process and what conditioning factors the biology literature recommends?
- Which model is good for estimation and prediction of disability and mortality probabilities and the effects of the sequential conditioning covariates on them?
  (1) a statistical multistate Cox regression model,
  (2) a feedforward multilayer perceptron (MLP) type of neural network model, or
  (3) a long short memory (LSTM) recurrent neural network (RNN) model?
- Three models are formulated as finite-state stochastic processes, with states as health states {normal health, diseased, disability, death}—disability that qualifies for a public insurance program.
- I describe the statistical multistate Cox regression model and the recurrent neural network (RNN) model. I point out the limitations and strength of each approach.
- I use the HRS data and take time-varying conditioning covariates to be demographics, education, biomarkers (such as BMI, CES-D, cognition) and health related behaviors such as smoking, and moderately vigorous exercising
- I use R and SAS for Statistics and Keras and Tensorflow 2.0 and Python for machine learning analysis.

Literature

Biomedical Mechanism of Health Developments

Genetic: Barondes (1999); Khoury et al. (2009); Bookman et al. (2011) Telomere: Hayflick (1965); Austad and Fischer (2016); Shalev and Belsky (2016); Simons et al. (2016) Epigenetic: Esteller (2008); Hannum et al. (2013) Programming and important milestones: Barker (1990); Barker (2007); Gluckman et al. (2008); Thornburg et al. (2010); Kanherkar et al. (2014); Simons et al. (2016)

Statistical models and estimation methods

Multistate statistical models: Aalen and Johansen (1978); Andersen, Borgan, et al. (1993); Andersen and Perme (2008); Crowther and Lambert (2017); Fleming (1978)

Machine Learning Models

MLP: universal function approximator, Hornik et al., 1989, applied to survival analysis: Faraggi and Simon, 1995; Katzman et al., 2018; Lee et al., 2018; Ranganath et al., 2016
LSTM-RNN:sequence-to-sequence map approximator, Graves et al., 2014; Hammer, 2000; Siegelmann and Sontag, 1992 application single event type Ren et al., 2019, and multi-events this paper.

The Basic Model: the finite state stochastic process

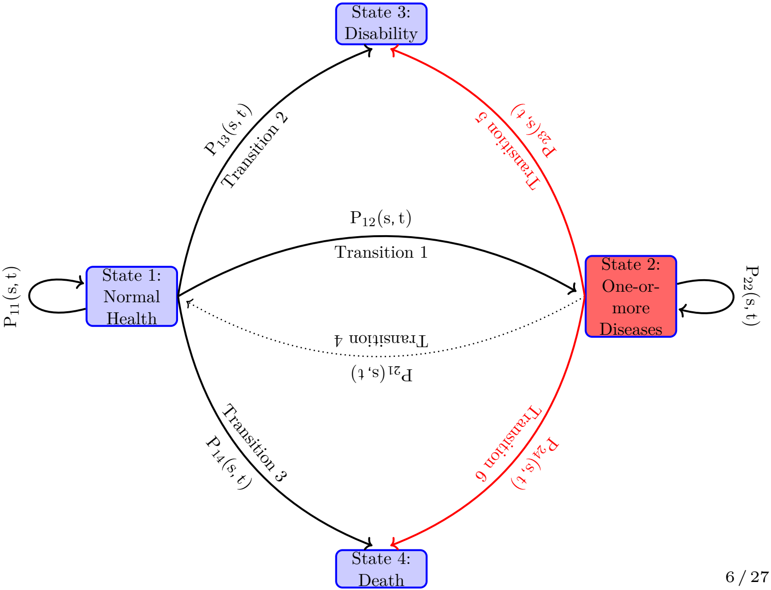
Assume that health outcomes over the life-cycle follow a multi-state continuous time Markov process $X(t), t \in T$ with the finite state space $S = \{1, ..., 4\}$.
$T = [51, 65]$ and unit of time is a year.
1 = Healthy or normal health,
2 = Diseased with one or more chronic diseases,
3 = Disability qualifying for the DI or SSI programs, and
4 = Death.
The transition probabilities of the stochastic process $X(t)$

$$P_{hj}(s,t) = \text{Prob}(X(t) = j | X(s) = h), \tag{1}$$

$h, j \in S$, $s, t \in T$, $t \geq s$.
Denote the matrix of transition probabilities by

$$P(s,t) \equiv \left(P_{hj}(s,t)\right)_{h,j=1...p}. \tag{2}$$

Path Diagram of health evolution over the life-span



Solving Transition Probabilities $P(s,t)$ non-parametrically

Repeated use of the Chapman-Kolmogorov equation on a subdivision $s = t_0 < t_1 < \ldots < t_m = t$:

$$P(s,t) = P(t_0,t_1) \cdot P(t_1,t_2) \cdot \ldots \cdot P(t_{m-1},t_m) = \prod_{i=1}^{m} P(t_{i-1},t_i) \tag{3}$$

Define transition intensity a.k.a hazard rate by

$$\lambda_{hj}(t) = \lim_{\Delta t \to 0} \frac{P_{hj}(t,t+\Delta t)}{\Delta t} \approx \frac{d_{hj}}{\bar{Y}_h} = \hat{\lambda}_{hj}(t). \tag{4}$$

Integrated transition intensity function for transition $h \to j$:

$$\Lambda_{hj}(t) = \int_0^t \lambda_{hj}(u)\, du \approx \sum_{t_{hj} \le t} \frac{d_{hj}}{\bar{Y}_h} = \hat{\Lambda}_{hj}(t). \tag{5}$$

Product Integral result: The transition probabilities of a stochastic process parameterized via an intensity process is given by the product integral of the integrated hazard function.

$$P(s,t) = \prod_{s}^{t} (I + d\Lambda(u)) \approx \prod_{s}^{t} \left(I + d\hat{\Lambda}(u)\right). \tag{6}$$

Econometric Specification and Estimation

Parameterization of transition intensities, for $h \to j$

$$\lambda_{hj}\left(t; X_h(t), \beta_{hj}\right) = \lambda_{hj,0}(t)\,\mathsf{exp}^{\beta'_{hj} X_h(t)} \tag{7}$$

Likelihood function of the data

- $Y_{h,i}(t) = 1$ if i is in health state h, $h = 1, 2$ at time t.
- $T^*_{h,i}$ = completed duration in state h or censoring time
- $N_{hj,i}(t) = \#$ of transitions of type $h \to j$ by i during $[0, t]$.
- Likelihood of the sample

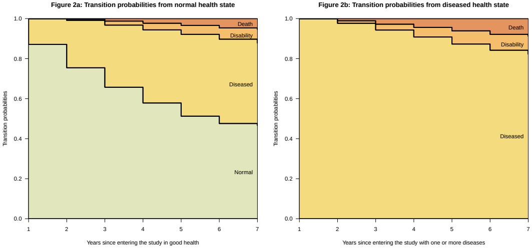
$$L(\theta) = \prod_i \prod_{\substack{h=1,2\\ j=2,3,4\\ h\neq j}} \left(\prod_t \lambda_{hj,i}\left(t|X_{h,i}\right)^{\triangle N_{hj,i}(t)}\right) \mathsf{exp}\left(-\int_0^{T^*_{h,i}} \lambda_{hj,i}\left(u|X_{h,i}\right)\right) du \tag{8}$$

Definition of variables

White and Female: Definitions are standard.

- College: Completed college or higher education (1) or not (0).
- cesd: Based on the score on the Center for Epidemiologic Studies Depression (CESD) measure, capturing the level of stress and depression.
- cogtot: A measure of cognitive functioning based on the aggregate scores on word recall, counting backwards, naming tasks (e.g., date-naming), and vocabulary questions.
- bmi: The standard body-mass-index (BMI).
- behav_smoke: A binary variable taking value 1 if the respondent ever smoked and 0 otherwise.
- behav_vigex: This variable takes value 1 if the respondent did vigorous exercises 3 or more times per week.

Transition probabilities: without covariates



duration	$1 \to 1$	$2 \to 2$	$1 \to 2$	$1 \to 3$	$2 \to 3$	$1 \to 4$	$2 \to 4$
0	1.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1	0.8711	1.0000	0.1289	0.0000	0.0000	0.0000	0.0000
2	0.7543	0.9767	0.2367	0.0062	0.0127	0.0027	0.0106
3	0.6570	0.9430	0.3105	0.0201	0.0292	0.0123	0.0279
4	0.5786	0.9082	0.3651	0.0331	0.0477	0.0232	0.0441
5	0.5126	0.8730	0.4087	0.0446	0.0659	0.0341	0.0610
6	0.4761	0.8421	0.4212	0.0562	0.0792	0.0465	0.0787
7	0.4667	0.8229	0.4116	0.0681	0.0896	0.0536	0.0874

Source: The author.

Estimates of Cox regression models separately for each transition with health measures

	1->2	1->3	1->4	2->3	2->4
cesd	0.5612***	1.9802***	0.1265	1.2950***	0.5467*
	(0.1095)	(0.3460)	(1.0202)	(0.1587)	(0.2461)
bmi	0.0423***	−0.0051	0.0118	0.0250**	−0.0213
	(0.0055)	(0.0313)	(0.0499)	(0.0088)	(0.0168)
cogtot	−0.0029	−0.0671**	0.0152	−0.0353***	−0.0091
	(0.0058)	(0.0256)	(0.0407)	(0.0094)	(0.0157)
behav_smoke	0.0454	0.2577	2.5107*	0.3814***	0.8173***
	(0.0508)	(0.2228)	(1.0166)	(0.1078)	(0.1777)
behav_vigex	−0.1966**	−0.9995***	−1.1687*	−0.6103***	−1.1988***
	(0.0712)	(0.2438)	(0.4997)	(0.1039)	(0.1431)
White	0.0452	0.0982	−0.4942	−0.1172	−0.3252*
	(0.0685)	(0.2823)	(0.5103)	(0.1080)	(0.1576)
College	−0.1112	−0.8871*	−1.1140	−0.6839***	−0.7443**
	(0.0648)	(0.4167)	(0.7564)	(0.1937)	(0.2571)
Female	0.0851	−0.2734	−0.6858	−0.0801	−0.2518
	(0.0515)	(0.2335)	(0.4694)	(0.1006)	(0.1445)
AIC	23311.3089	1358.1202	318.1039	7268.4078	3268.0150
R^2	0.0294	0.0252	0.0084	0.0366	0.0217
Max. R^2	0.9993	0.3483	0.0961	0.7029	0.4349
Num. events	1500	95	23	446	207
Num. obs.	3239	3334	3262	6165	5926
Missings	344	361	390	691	929
PH test	0.0237	0.2294	0.0447	0.0001	0.1660

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

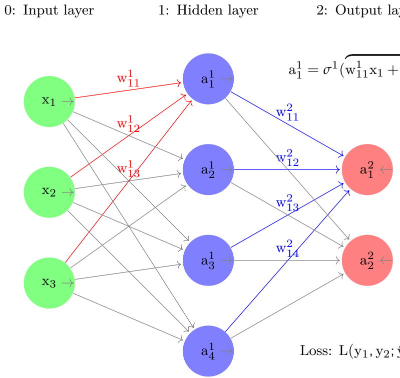
MLP Architecture

$$\hat{y} = f(x; w) \equiv f^{L}_{w^L} \circ .... \circ f^{1}_{w_1}(x). \tag{9}$$

$$\min_{w} L(y, f(x, w)) + \lambda C(w). \tag{10}$$

Special Case: $L = 2, x \in \Re^3, y \in \Re^2$, # hidden layer neurons $= 3$.

$$\hat{y} = f(x; w) = \sigma^2\left(z^2\left(\sigma^1\left(z^1(x, w^1)\right), w^2\right)\right) \equiv f^2_{w^2} \circ f^1_{w^1}(x).$$



$$a^1_1 = \sigma^1(\overbrace{w^1_{11}x_1 + w^1_{12}x_2 + w^1_{13}x_3 + b^1_1}^{z^1_1})$$

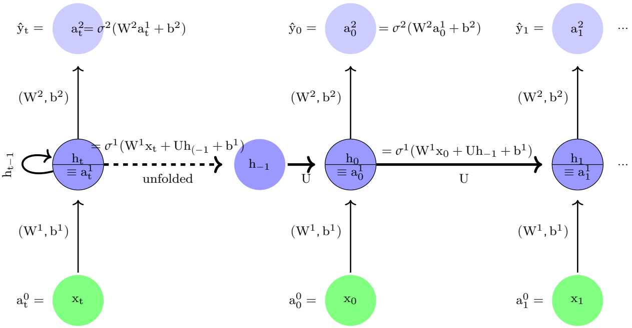
$$a^2_1 = \sigma^2(\overbrace{\sum_{j=1}^{4} w^2_{1j}a^1_j + b^2_1}^{z^2_1}) \equiv \hat{y}_1$$

$$a^2_2 = \sigma^2(\overbrace{\sum_{j=1}^{4} w^2_{2j}a^1_j + b^2_2}^{z^2_2}) \equiv \hat{y}_2$$

Loss: $L(y_1, y_2; \hat{y}_1, \hat{y}_2)$; $L = 2$ is the last layer

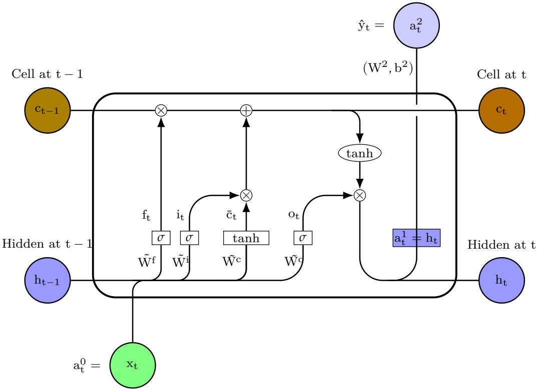
RNN Architecture

$$
\begin{aligned}
h_t &= \sigma^1(W^1 \cdot x_t + U \cdot h_{t-1} + b^1) \\
a_t^1 &\equiv h_t \\
a_t^2 &= \sigma^2(W^2 \cdot a_t^1 + b^\ell) \\
\hat{y}_t &\equiv a_t^2 \text{ (notational convenience)}
\end{aligned}
$$



An LSTM memory cell at time-step t

$$
\begin{aligned}
f_t &= \sigma\left(W^f \cdot x_t + U^f \cdot h_{t-1} + b^f\right) \\
c_t &= f_t * c_{t-1} + i_t * \bar{c}_t \\
h_t &= o_t * \tanh(c_t)
\end{aligned}
$$



Functional specifications of an LSTM memory cell

Layer 1 (hidden recurrent layer):

$$
\begin{aligned}
\mathrm{f_t} &= \sigma\left(\mathrm{W^f}\cdot \mathrm{x_t} + \mathrm{U^f}\cdot \mathrm{h_{t-1}} + \mathrm{b^f}\right) \\
\mathrm{i_t} &= \sigma\left(\mathrm{W^i}\cdot \mathrm{x_t} + \mathrm{U^i}\cdot \mathrm{h_{t-1}} + \mathrm{b^i}\right) \\
\mathrm{o_t} &= \sigma\left(\mathrm{W^o}\cdot \mathrm{x_t} + \mathrm{U^o}\cdot \mathrm{h_{t-1}} + \mathrm{b^o}\right) \\
\bar{\mathrm{c}}_\mathrm{t} &= \tanh(\mathrm{W^c}\cdot \mathrm{x_t} + \mathrm{U^c}\cdot \mathrm{h_{t-1}} + \mathrm{b^c}) \\
\mathrm{c_t} &= \mathrm{f_t} * \mathrm{c_{t-1}} + \mathrm{i_t} * \bar{\mathrm{c}}_\mathrm{t} \\
\mathrm{h_t} &= \mathrm{o_t} * \tanh(\mathrm{c_t}) \\
\mathrm{a_t^1} &\equiv \mathrm{h_t} \\
\mathrm{h_{-1}}, \mathrm{c_{-1}} &= \text{ user supplied initial activation levels}
\end{aligned} \tag{11}
$$

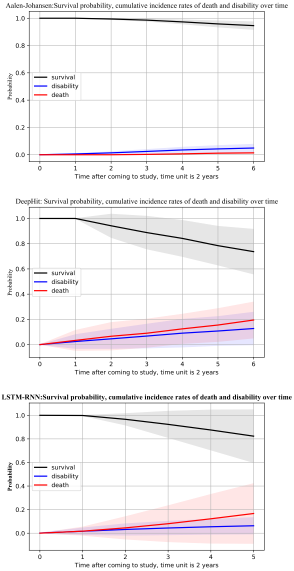
Layer 2 (Output): This is same as the RNN output layer.

$$
\begin{aligned}
\mathrm{a_t^2} &= \sigma^2(\mathrm{W^2}\cdot \mathrm{a_t^1} + \mathrm{b^\ell}) \\
\hat{\mathrm{y}}_\mathrm{t} &\equiv \mathrm{a_t^2} \text{ (notational convenience)}
\end{aligned} \tag{12}
$$

Average predicted cumulative incidence rates of disability and death in the test sample

up to time	Statistical multistate model		Lee etal Deephit model		LSTM-RNN model	
	Disability	Death	Disability	Death	Disability	Death
0	0.00000	0.00000	0.00000	0.00000	0.00119	0.00028
1	0.00573	0.00000	0.03487	0.00001	0.01658	0.01725
2	0.01383	0.00000	0.05179	0.00396	0.03165	0.04515
3	0.02420	0.00250	0.12043	0.02575	0.04405	0.08047
4	0.03489	0.00666	0.24674	0.04990	0.05461	0.12219
5	0.04276	0.01136	0.27139	0.13317	0.06310	0.16697
6	0.04906	0.01379	0.28676	0.15479	0.07025	0.20781
7	0.05171	0.01705	0.49585	0.50415	0.07734	0.24754
c-index	0.476290706		0.74824416		0.755676489	

Mean and confidence intervals of predicted probabilities on test data



Conclusions

- Judging from the c-index estimates on the test data, the widely used Aalen-Johansen-Cox type statistical estimation method in the multi-state framework performs much worse than the neural network models in predicting the transition probabilities. It is important to extend the linearity and proportionality assumption for the statistical model.
- The sequential LSTM-RNN model is more appropriate than the feedforward MLP type neural networks. BUT the neural network models have disadvantages of overfitting and being not readily able to estimate the effects of covariates. Need more work along this line.
- From the parameter estimates of the statistical model, the paper finds the most important factors to be — CES-D measuring level of depression and stress, with positive(i.e. unfavorable) effects and college education, with negative (favorable) effects on almost all health transitions.
- Other important factors: Smoking with significant adverse effect and exercising regularly with favorable effects on most health outcomes.

The End

Aalen, O. O. and S. Johansen (1978). An Empirical Transition Matrix for Non-Homogeneous Markov Chains Based on Censored Observations. Scandinavian Journal of Statistics, 5, no. 3, 141–150 (cit. on p. 4).

Andersen, P. K., Ø. Borgan, R. D. Gill, and N. Keiding (1993). Statistical Models Based on Counting Processes. Springer-Verlag, New York (cit. on p. 4).

Andersen, P. K. and M. P. Perme (Sept. 2008). Inference for outcome probabilities in multi-state models. Lifetime Data Analysis, 14, no. 4, 405–431 (cit. on p. 4).

Austad, S. N. and K. E. Fischer (2016). Sex Differences in Lifespan. Cell Metabolism, 23, no. 6, 1022–1033 (cit. on p. 4).

Barker, D. J. P. (May 2007). The origins of the developmental origins theory. Journal of Internal Medicine, 261, no. 5, 412–417 (cit. on p. 4).

Barker, D. J. P. (1990). The fetal and infant origins of adult disease. BMJ: British Medical Journal, 301, no. 6761, 1111 (cit. on p. 4).

Barondes, S. (1999). Molecules and Mental Illness. Scientific American Library (cit. on p. 4).

Bookman, E. B., K. McAllister, E. Gillanders, K. Wanke, D. Balshaw, J. Rutter, J. Reedy, D. Shaughnessy, T. Agurs-Collins, D. Paltoo, and et al. (2011). Gene-environment interplay in common complex diseases: forging an integrative model-recommendations from an

Multistate statistical Model vs LSTM-RNN

Lakshmi K. Raut

NIH workshop. Genetic Epidemiology, n/aâ€“n/a (cit. on p. 4).

Crowther, M. J. and P. C. Lambert (2017). Parametric multistate survival models: Flexible modelling allowing transition-specific distributions with application to estimating clinically useful measures of effect differences. Statistics in Medicine, 36, no. 29, 4719–4742 (cit. on p. 4).

Esteller, M. (Mar. 2008). Epigenetics in Cancer. New England Journal of Medicine, 358, no. 11, 1148–1159 (cit. on p. 4).

Faraggi, D. and R. Simon (Jan. 1995). A neural network model for survival data. Statistics in Medicine, 14, no. 1, 73–82 (cit. on p. 4).

Fleming, T. R. (1978). Nonparametric Estimation for Nonhomogeneous Markov Processes in the Problem of Competing Risks. The Annals of Statistics, 6, no. 5, 1057–1070 (cit. on p. 4).

Gluckman, P. D., M. A. Hanson, C. Cooper, and K. L. Thornburg (July 2008). Effect of In Utero and Early-Life Conditions on Adult Health and Disease. New England Journal of Medicine, 359, no. 1, 61–73 (cit. on p. 4).

Graves, A., G. Wayne, and I. Danihelka (2014). Neural Turing Machines. CoRR, abs/1410.5401 (cit. on p. 4).

Hammer, B. (2000). On the approximation capability of recurrent neural networks. Neurocomputing, 31, no. 1, 107–123 (cit. on p. 4).

Hannum, G. et al. (2013). Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates. Molecular Cell, 49, no. 2, 359–367 (cit. on p. 4).

Hayflick, L. (1965). The limited in vitro lifetime of human diploid cell strains. Experimental Cell Research, 37, no. 3, 614–636 (cit. on p. 4).

Hornik, K., M. Stinchcombe, and H. White (Jan. 1989). Multilayer feedforward networks are universal approximators. Neural Networks, 2, no. 5, 359â€“366 (cit. on p. 4).

Kanherkar, R. R., N. Bhatia-Dey, and A. B. Csoka (Sept. 2014). Epigenetics across the human lifespan. Frontiers in Cell and Developmental Biology, 2, no. 3, 124–128 (cit. on p. 4).

Katzman, J. L., U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger (Feb. 2018). DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Medical Research Methodology, 18, no. 1, 24 (cit. on p. 4).

Khoury, M. J. et al. (June 2009). Genome-Wide Association Studies, Field Synopses, and the Development of the Knowledge Base on Genetic Variation and Human Diseases. American Journal of Epidemiology, 170, no. 3, 269–279 (cit. on p. 4).

Lee, C., W. R. Zame, J. Yoon, and M. van der Schaar (2018). "Deephit: A deep learning approach to survival analysis with competing risks". Thirty-Second AAAI Conference on Artificial Intelligence (cit. on p. 4).

Ranganath, R., A. Perotte, N. Elhadad, and D. Blei (2016). "Deep Survival Analysis". Proceedings of the 1st Machine Learning for Healthcare Conference. Vol. 56. PMLR, 101–114 (cit. on p. 4).

Ren, K., J. Qin, L. Zheng, Z. Yang, W. Zhang, L. Qiu, and Y. Yu (2019). "Deep recurrent survival analysis". Proc. AAAI, 1–8 (cit. on p. 4).

Shalev, I. and J. Belsky (2016). Early-life stress and reproductive cost: A two-hit developmental model of accelerated aging?, Medical Hypotheses, 90, 41–47 (cit. on p. 4).

Siegelmann, H. T. and E. D. Sontag (1992). "On the computational power of neural nets". Proceedings of the fifth annual workshop on Computational learning theory - COLT '92. ACM Press (cit. on p. 4).

Simons, R. L., M. K. Lei, S. R. Beach, R. A. Philibert, C. E. Cutrona, F. X. Gibbons, and A. Barr (2016). Economic hardship and biological weathering: The epigenetics of aging in a U.S. sample of black women. Social Science & Medicine, 150, 192–200 (cit. on p. 4).

Thornburg, K. L., J. Shannon, P. Thuillier, and M. S. Turker (2010). "In Utero Life and Epigenetic Predisposition for Disease". Epigenetics and Cancer, Part B. Elsevier, 57–78 (cit. on p. 4).