Biomarker Adoption in Developmental Science: A Data-driven Modeling of Popularity Trends from 90 Biomarkers Across Time

Weiqiang Qian[1], Chao Zhang[1], Hannah A. Piersiak[1], Kathryn L. Humphreys[1]*, & Colter Mitchell[2]*

[1]Vanderbilt University

[2]University of Michigan

*Shared senior co-authorship

Short Running Title: Biomarker Use in Developmental Science

Corresponding authors. Kathryn L. Humphreys (k.humphreys@vanderbilt.edu) & Colter Mitchell (cmsm@umuch.edu)

Data, materials and code availability statement: The data that support the findings of this study are available at https://osf.io/ksj2n/?view_only=5eea3aeaf3b541dba289f226c026c63e, DOI 10.17605/OSF.IO/KSJ2N.

Conflict of interest disclosure: The authors declare no conflicts of interest.

Acknowledgements: We thank Candice Odgers, Terrie Moffit, and Avshalom Caspi for their assistance in conceptualizing this project. This work was supported by the Jacobs Foundation (2019-1317-00 to K.L.H. and C.M.). The authors declare no competing financial interests. Dr. Humphreys was also supported in part by the National Science Foundation (2042285). Dr. Mitchell was also supported in part by the National Institutes of Health R01 AG071071 & R01 MD011716.

Author Contribution Statement

Weiqiang Qian: Formal analysis (supporting); methodology (lead); software (equal) writing – original draft (lead); writing – review and editing (equal). **Chao Zhang:** Formal analysis (lead); methodology (supporting); software (equal). **Hannah Piersiak:** Methodology (supporting); writing – original draft (supporting); review and editing (equal). **Kathryn Humphreys:** Conceptualization (supporting); funding acquisition (supporting); methodology (supporting); supervision (supporting); writing – original draft (equal); writing – review and editing (equal); methodology (supporting); supervision (lead). **Colter Mitchell:** Conceptualization (lead); funding acquisition (lead); methodology (supporting); supervision (supporting); writing – original draft (equal); writing – review and editing (equal);

RUNNING HEAD: BIOMARKER ADOPTION

Biomarker Adoption in Developmental Science: A Data-driven Modeling of Popularity Trends from 90 Biomarkers Across Time

Abstract

Developmental scientists have largely adopted numerous biomarkers in their research to better understand influences of adversity on development and risk for long-term health outcomes. Yet, the patterns of adoption and ultimate utility of biomarkers in developmental research merits investigation given the substantial time and financial resources used to include biomarkers in developmental research with infants and children. In the present paper we document trends of use of 90 biomarkers between 2000–2020 from approximately 430,000 publications indexed by the Web of Science. We used a data-driven approach to estimate biomarker growth trajectory based on yearly publication number, NIH dedicated funding resources, journal impact, years since the first publication, and number of research institutions involved in the biomarker research. Results indicate that most biomarkers experience rapid growth followed by a plateau. External funding, number of research institutions involved, and variability of journal impact all are associated with variations in growth.

Keywords: metascience; biomarkers; developmental science; developmental psychology

Biomarker Adoption in Developmental Science: A Data-driven Modeling of Popularity Trends from 90 Biomarkers Across Time

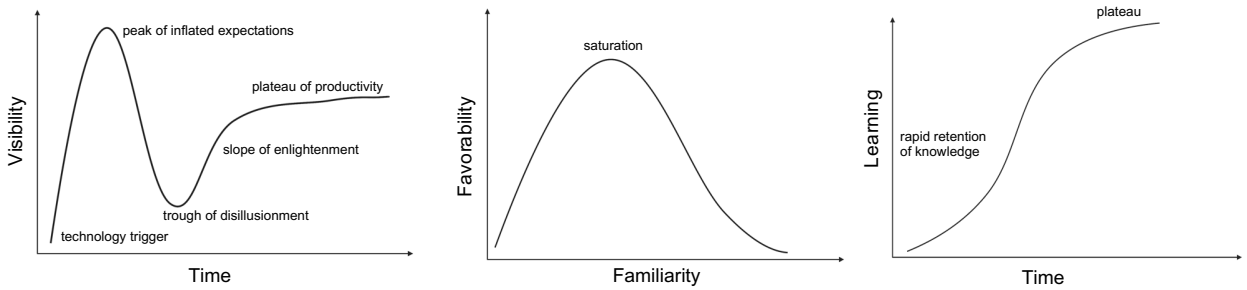
The integration of biological data into child development research, while not new, has rapidly increased over the last two decades, from measures of hormone levels to brain structure to molecular measures such as DNA and epigenetics. Scholars who study child and human development have been trying to integrate biological theory, methods, and data for centuries (Bernard, 2012; Meloni, 2014, 2016). Early theories of social processes were squarely rooted in Darwin's theory of evolution, but even before that, societies were thought to have life cycles like many mammals. Often biology provides great insights into social processes, but it has also run amuck with disastrous effect (e.g., phrenology, eugenics; Meloni, 2016; Thornton, 2005). However, since the late 20th century there has been a particularly rapid integration of measures of biology into child development research. Several rationales have been provided for integrating biomarkers with social and behavioral data (Falk et al., 2013; Meloni, 2016; Mitchell et al., 2013). For most developmental outcomes, biology is expected to be a mechanism linking social (and other) exposures to development. Also, of particular interest to developmental scientists, biomarkers could be used as indicators of exposure so that even if not part of the causal pathway from external stimuli to outcome, researchers can reconstruct the (sometimes unobserved) past. Similarly, many biomarkers were developed to indicate disease before typical symptoms emerge (Falk et al., 2013; Justice et al., 2018; Martin et al., 2019; Mitchell, Schneper, & Notterman, 2016; Singh & Rose, 2009). For others even if the biomarker was not part of the causal pathway from external stimuli to outcome, the consideration of biomarkers may improve the estimate of the variable of interest on development (i.e., controls). In addition to these more scientific reasons, others have suggested that the prominence of biology as a scientific field has led to its integration into less prestigious areas of science to improve scientific reputation of practitioners (Bernard, 2012).

Yet, most biomarkers were not developed with the goal of examining variation related to social behavior, to study mechanisms linking environment to child development outcomes, or to be measured outside of a lab or clinic. Furthermore, most were not developed with children as the target (Justice et al., 2018; Meloni, 2016; Mitchell et al., 2016; Singh & Rose, 2009). Attempts to identify biomarkers relevant to developmental stage, consequences of adversity, and risk for negative health outcomes have been widely embraced by developmental scientists. However, whether a biomarker will be useful is often difficult to determine, particularly for scholars with no or limited advanced training in biology. Small samples (see Davis-Kean & Ellis, 2019; Ritchie, 2020) and incorrect interpretation of biomarkers (for an example of this discussion with DNA methylation, see Moore & Kobor, 2020) are regularly observed as limits to accuracy in most studies of infants and children, and likely contribute to a replication crisis. Unfortunately, despite good intentions, developmental scientists may be participants in a problematic cycle of booming popularity and decreasing quality, with significant consequences for precious scientific resources (e.g., funding, time, supplies). Importantly, biomarkers also often fail to live up to their initial promise.

In this paper we document the patterns of adoption and use of biological metrics in relation to child development research. Although there is nothing theorized specifically about models of biomarker integration into developmental sciences, others have speculated about the effect of novel findings and technical innovations (i.e., new data, new statistical technique, unique finding) on a field or market (Fenn, 1995; Fenn & Linden, 2003; O'Leary, 2008; Ritchie, 2020). For example, innovative technologies are thought to emerge and change as part of a hype curve or cycle. This hype curve is thought to have five general stages (see Figure 1A; Fenn & Raskino, 2008): 1) the technology trigger phase when the technology growth is driven by the potential of the innovation, 2) the peak of inflated expectations phase when several initial success stories are publicized but access to the technology is still limited, 3) the trough of disillusionment when the hype rapidly declines due to realistic re-adjustment of the innovation's

potential impact, 4) the slope of enlightenment, a long phase in which an often small group of cautious users remain attentive to how the innovation is appropriately used and 5) the plateau of productivity in which the true advancement into the broader field can take place. More broadly though, social and developmental sciences have had a long history of also attempting to examine how scientific methods and ideas spread through a field. Two very prominent social psychological theories that might apply are the infatuation/novelty curve and the learning curve. The infatuation or novelty curve (see Figure 1B) was first proposed in 1968 and generally suggests that as familiarly grows so does favorability until saturation is reached and then begins to decline (Berlyne, 1970). The learning curve (see Figure 1C) is even older (1924) and is essentially a cumulative density function related to knowledge such that more rapid retention of knowledge occurs at the beginning, slowing over time until it becomes flat toward the end (Adler & Clark, 1991; Baloff, 1966). Some have suggested that the innovative technology hype curve is just the combination of an infatuation curve with a slower learning curve following it so that the real growth comes after a bout of infatuation (Dedehayir & Steinert, 2016). It is this potential insight into the underlying processes that will afford us to better quantify the underlying processes.

Figure 1. Potential Patterns of Adoption and Use



A. The Hype Curve B. The Infatuation/Novelty Curve C. The Learning Curve

Note. Figure created with BioRender.com based on prior research (Adler & Clark, 1991; Baloff, 1966; Berlyne, 1970; Fenn & Raskino, 2008).

The degree to which biomarkers follow a specific trend (hype curve, infatuation curve, learning curve), allows us to document the growth in biomarker use and make inferences about patterns of adoption. Further, this approach provides the opportunity to make predictions on future use of biomarkers. Our goals are to: (1) document historic patterns of biomarker use, and (2) utilize those observations to offer a generalizable predictive model that can be applied to foreshadow future trends. To do this, we gathered publication related statistics for 90 biomarkers to create a set of content agnostic predictive models that both explain and predict publication popularity.

Method

Biomarker Selection

In our modeling plan, each biomarker is the modeling unit in analyses. Thus, our first step was to identify a candidate list of biomarkers used in developmental science by searching consortia, large data sets, and databases (not just developmental) such as the USC biomarker network, National Study of Adolescent to Adult Health, Environmental Influences on Child Health Outcomes, and meta-analyses (e.g., Justice et al., 2018). Although this was not a list of all potential biomarkers, it was intended to be large enough to derive sufficient data to generate a predictive algorithm. A round of screening on the original list of biomarkers was conducted based on the availability of developmental publications in the Web of Science and NIH funding. Biomarkers with too few publication entries or NIH-funded projects were not retained for further analysis, leaving 90 biomarkers for the next stage of analysis. For the list of biomarkers, and the number of publications and NIH funded projects, please see Supplemental Table 1.

Variables of Interest

The statistics of interest relevant to each biomarker include the number of publications, the number of projects funded by National Institute of Health (NIH), the amount of funding given by NIH, the number of institutes participating in publication activities, and the quality of

publication outlets. All of the statistics included can be obtained from publicly available sources with relative ease. These statistics are by no means the only ones we think are relevant to predicting the research popularity. Nevertheless, they capture important behavioral information related to the publication activities as would be explained.

Yearly Publication Number

The yearly number of publications is the most important variable in our modeling process. Because interest in biomarkers will be largely translated to the number of publications, it is logical to use publication number as the outcome variable of our modeling process. In addition to serving as the outcome variable, the publication numbers from previous years serve as important indicators to current research interest.

We used Web of Science to query publication information for each of the selected biomarkers. Web of Science offers access to scholarly work from a broad range of scientific disciplines, allowing us to capture the developmental science adoptions of biomarkers with specific emphasis such as medical or social applications. For each biomarker, we used the following statement to query relevant publications within developmental science: “TS = (the query for each biomarker) AND (CHILD* OR ADOLESCENT* OR INFANT* OR NEWBORN* OR PEDIATRICS*)”. All biomarker names were quoted in the search, to avoid partial matches to the biomarker names. All entries returned from the query were retained. Publication data after 2020 were removed due to incompleteness. Publications prior to 2000 were also removed, as we planned to focus on the time range where most of the included biomarkers have adequate publication data coverage across the modeling timespan. Between 2000 and 2020, among the 90 biomarkers, the number of publications ranged from 0 to 4635, with a median of 90 publications.

Considering changes in publication numbers across time requires including the potential influence of inflation in article publications. As seen in the Dimension database that tracks the overall publication numbers across years, the yearly number of all scientific publications

increases over time (Bode, Herzog, Hook, & McGrath, 2018). To make the number of publications comparable across years so we could model the ebb and flow in the research interest for each biomarker, the queried publication number for biomarkers in a given year was divided by the total scientific publication number of that year from the Dimension database.

Publication Growth Rate

The growth rate of publications related to each biomarker for a given year captures the acceleration in publication interest. The rate of change for each year was defined to be the percentage change in the rescaled publication number relative to the statistic from the previous year. Between 2000 and 2020, among the 90 biomarkers, the growth rate ranged between 89.2% and 966.8%, with a median of 1.6%.

Participating Institutes

The number of institutes involved in publications for each biomarker offers information on the penetration of a biomarker into the field. The number of institutes interested in a particular biomarker is largely unrestricted except for the total number of existing research institutes, especially for biomarkers not requiring expensive and specialized assaying processes, thus allowing for rapid growth of research output related to particular biomarkers.

We obtained the yearly total number of involved research institutes by tabulating the author affiliations from each of the publications gathered for our analysis from Web of Science. We counted the number of unique institutes found from all publications related to a particular biomarker from a particular year. Between 2000 and 2020, among the 90 biomarkers, the yearly affiliation number ranged between 0 and 5962, with a median of 188 institutes per biomarker per year.

NIH Funding

Funding provides critical support to research activities, especially for biomarker research that typically involves expensive assaying processes. We are interested in both the total funding amount associated with each biomarker, as well as the number of funded projects. While the

total funding amount made available to specific biomarkers is an indicator of the resources supporting the research popularity, it may be accounted for by only a few large grants. Thus, we also included the number of projects funded to capture the spread of funding support that drives the popularity for specific biomarkers.

The number of NIH projects funded, as well as the monetary amount of funding from NIH were used to approximate funding availability for each biomarker. We selected NIH due to the wide range of scientific disciplines it supports, as well as the number of grants it awards. NIH funding information of relevant projects was queried in the Research Portfolio Online Reporting Tools (RePORT) on the NIH website. The following query for NIH-funded projects was used: “Text Search: (the query for each biomarker) and (child or children or infant or newborn or pediatrics) (advanced)”. The projects with the same project ID were merged in each year, with their total funding summed. Some projects had sub-projects, which created overlaps with parent projects. To solve this problem, only sub-projects whose parents were not included in data were retained. The searches were conducted on December 31, 2020. All data after 2020 were removed due to incompleteness. Between 2000 and 2020, among the 90 biomarkers, the number of funded projects ranged from 0 to 2186, with a median of 21 projects funded by NIH.

Similar to publication increases, both the number of projects funded and the total amount of funds from NIH have seen increases across years, thus requiring adjustment to make the research resources available to biomarkers comparable across years. To do so, we obtained the yearly total number of projects funded by NIH, as well as the yearly sum of funding provided to all projects, to serve as denominators in the corresponding adjustments. The yearly total statistics were obtained from RePORT through querying all projects from a given year without supplying any search terms.

Ranking of Publication Outlets

The increase in popularity for some biomarkers could be partially attributed to publications in journals with lower impact factors, as more researchers compete for similar work

to be published in higher impact outlets. While journal impact does not necessarily indicate something meaningful about the quality of journals or a given article published in that journal, this metric nevertheless reflects one method for assessing prestige within the scientific community. By including journal rankings to predict the popularity, we tested if rapidly rising popularity co-occurs with a higher proportion of lower ranked journal publications.

To evaluate the impact of journals, we used the SCImago Journal Rank indicator (SJR indicator). The SJR indicator assesses the impact of journals contained in the Scopus database yearly (Scimago Lab). The SJR indicator is available for journals between 1999 and 2020.

Each of the publications queried from our analysis was given a SJR indicator value through matching the journal name from Web of Science query to the journals in the SJR database. However, some publications cannot be given a SJR value because some journals queried from Web of Science were not included in SJR in specific years. To solve this problem, a 0 was given as the SJR indicator for these publications, with the assumption that journals not included in the SJR database may not have established high impact, such as journals that were newly established.

For our modeling approach, we used both the yearly average SJR across all publications on a specific biomarker, and the yearly variance of SJR indicator for publications on a biomarker. The yearly average captures the overall impact of biomarker research, whereas the variance captures the spread of the impact of specific biomarker research. Between 2000 and 2020, among the 90 biomarkers, the yearly average SJR ranged from 0 to 14302, with a median of 1272.; the yearly median SJR ranged from 0 to 14302, with a median of 966; and the yearly variance in SJR ranged from 868314 to 54036606, with a median of 1540092. SJR impact data was not rescaled in our analysis.

Time since first adoption

One final predictor was the number of years since a biomarker was first adopted in developmental science. Beyond understanding the general progression of popularity through

time, we can also better capture the effect of other predictors included through controlling the effect of time. Among the 90 biomarkers, the number of years since biomarker adoption ranged from 0 to 115, with a median of 27.

Modeling Frameworks

We propose two modeling frameworks, a polynomial multilevel model and a random forest model, to predict the yearly publication popularity several years into the future based on past publication statistics. The exact span of future prediction depends on the specific predictors used for the model. Both models used in our analysis were also able to identify the most important driving forces behind publication popularity, while the multilevel model can further elucidate the direction of relation between popularity and predictors.

Polynomial Multilevel Model

The polynomial multilevel model offers a flexible approach to model trajectories by accounting for the effects of time and other predictors on publication popularity, while also modeling the dependency among data from different years for the same biomarker. Because of the dependencies in data, treating yearly observations for the same biomarker as independent would violate assumptions for linear regressions, thus warranting the use of multilevel models.

In formulating the multilevel model, each biomarker was chosen as a level 2 unit, while the yearly publication data for biomarkers formed the level 1 observations. To avoid the issue of multicollinearity, we did not include all possible predictors from 3 to 5 years ago as fixed effect predictors. For fixed effects, both a forward selection and a backward selection procedure were used to retain predictors that were shown to have yielded the largest decrease in AIC and BIC at each step. Both selection processes stopped when adding or leaving out any of the variables no longer yielded lower AIC and BIC. The forward and the backward selection processes found the same set of variables, including linear, quadratic, and cubic number of years since the first publication, as well as affiliation number (5 years ago) and scaled NIH funded project number (5 years ago). This set of variables selected allows for making predictions 5 years into the future,

as the only historical statistics used are from 5 years ago. For random effects, the random intercept and the random slope of year number were used to represent the different size and the speed of increases in publication numbers for different biomarkers.

Because some of our variables can be tiny in terms of absolute quantity, such as the very small number for the adjusted yearly publication number, some of the model coefficients would be extremely small, thus hindering our interpretation. Accordingly, to avoid tiny coefficients in models while retaining the distributions of variables included, we used a min max normalization process, which rescaled all the variables within a range between 0 to 1. In the polynomial multilevel model, the rescaling was conducted across all biomarkers, rather than within each biomarker. The rationale was that the differences between magnitudes of data for biomarkers are readily accounted for by random intercepts and random slopes, leaving us to interpret the effect of predictors largely unaffected by the biomarker-specific effects.

Random Forest Regression

Random forest regression has been used widely in modeling highly nonlinear and interactive relations among many predictors and a continuous outcome with high accuracy. The model averages the predictions from many decision trees, each containing some split decision point at which a random set of input variables are entered to yield a single variable that reduces the prediction error in the outcome variable the most. Each tree is trained with a random subset of the data. The decision tree structure quite exhaustively learns the relation between input variables and output variable, thus risking overfitting the model to the specific training dataset. The design of the random forest model promotes the generalizability of the learned model to other datasets with both the random input variable selection at decision points and the random data sub-setting in the training process, all the while using an ensemble of trees to average the prediction output. For our purpose of predicting publication interest with many predictors, as well as offering a generalizable model to other biomarkers not included in our dataset, random forest regression provides a good fit.

Random forest models also allow for some level of interpretability through computing the contribution of each input variable to the reduction of outcome variable prediction error. The relative contribution for each of the included predictors is called feature importance. We planned to use feature importance to interpret the random forest model, as well as to guide the formulation of theory about publication popularity.

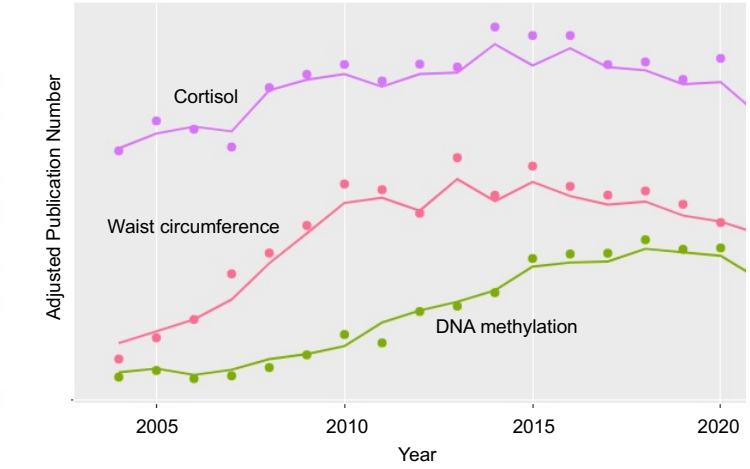
For the predictors, we included the 1st to 4th power of year number, as well as the following variables from 3 to 5 years ago: yearly NIH grant project number, yearly total NIH funding amount, yearly NIH mean funding amount per project, yearly affiliated institute number, yearly growth rate of publication number, yearly mean SJR indicator, and yearly variance of SJR indicator. The outcome variable was the adjusted yearly publication interest. To reduce the impact of differences among biomarkers, min max normalization was conducted on all covariates except year numbers and growth rates for each biomarker.

Results

Descriptive Data

In addition to providing a dataset of the 90 biomarkers for each of the 20 years (https://osf.io/ksj2n/?view_only=5eea3aeaf3b541dba289f226c026c63e), we created a visualization tool for charting the trends within developmental science research (https://hypecurve.shinyapps.io/hypecurve/). Broadly speaking, more novel biomarkers show an early rapid growth followed by a plateau that matches developmental publications overall (for an example of three biomarkers for illustrative purposes, see Figure 2).

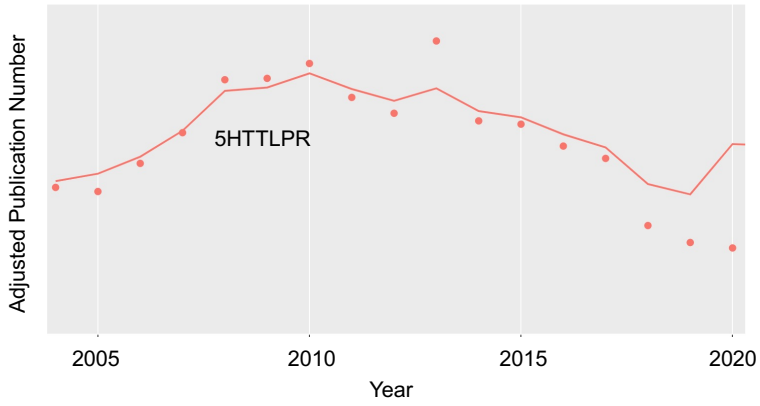
Figure 2. Growth of Three Illustrative Biomarkers in Developmental Sciences



Note. Dots represent values adjusted for total developmental publications; Lines represent modeled fit from the random forest model) for three illustrative biomarkers (i.e., cortisol, waist circumference, and DNA methylation).

Older biomarkers (e.g., cortisol) already have higher numbers of publications but slower growth rates as well. In general, the vast majority of biomarkers followed a similar trajectory of faster early growth (compared to developmental publications in general) followed by a slower change (typically a plateau or even small decline relative to the developmental literature in general). This suggests that most biomarkers follow a “learning curve” style growth. However, some biomarkers, such as *5-HTT* have growth curves that are very different (see Figure 3) and more akin to an infatuation or hype curve with rapid growth and similarly rapid decline in publications.

Figure 3. Growth of *5-HTT* in Developmental Sciences



Note. Dots represent values adjusted for total developmental publications; Line represents modeled fit from the random forest model).

Multilevel Model

The polynomial multilevel model was fitted with the lme4 package in R (v1.1-26; Bates, Maechler, Bolker, & Walker, 2015), with random slope for the linear effect of year number allowed for different biomarkers. All predictors included were on level-1, the level of yearly publication outcomes. No level-2 predictors were used in our analysis. The polynomial multilevel model was fitted with the whole dataset, to obtain parameter estimates and overall model performance in explaining variance with the fixed effect part of the model.

Table 1. Parameter estimates and significance information

	Estimates	Standard Error	df	T-value	Pr(>|t|)
Intercept	$-5.05x19^{-2}$	$3.08x10^{-2}$	130	-1.64	.103
Linear year number	$4.95x10^{-3}$	$1.37x10^{-3}$	337.5	3.61	<.001
Quadratic year number	$-6.17x10^{-5}$	$2.31x10^{-5}$	710.1	-2.67	.008
Cubic year number	$4.56x10^{-7}$	$1.14x10^{-7}$	633.4	4.02	<.001

Affiliations 5	-8.72×10^{-5}	5.11×10^{-6}	1296	-17.06	<.001
Adjusted NIH project number 5	5.60×10^{-1}	2.64×10^{-1}	1347	2.12	.034

Table 1 contains the parameter estimates and significance information. Besides the intercept, all retained predictors are significant. The linear and cubic year numbers are positively associated with the outcome variable, but the quadratic year number has a negative association. The coefficient for the year related predictors means that taken together, the initial linear increase is gradually overtaken by a decrease from the quadratic effect, while the cubic term, despite its very small coefficient, exerts a larger positive association with the passage of time. The affiliation number has a negative coefficient on publication popularity five years into future, whereas the NIH project number has a positive coefficient. Because the effect sizes of multilevel models is a topic that is still being researched, with no definitive approach to compute the effect sizes as would for the case of linear regression, we did not obtain effect sizes for predictors. The model fitted on the full dataset has an R^2 of .101 for the fixed effect. Similarly with the small magnitude of the outcome variable, the assessment of predictive ability relying on RMSE was no longer meaningfully interpretable. As a result, we took the absolute difference between the predicted value and the actual value, to compute a percentage of prediction deviation with the predicted value as the base. This statistic offered intuitive interpretations and comparability across different models. The predicted value was chosen as the base rather than the actual value because the possible actual value of 0 would return an infinity error. The median percentage deviation for the multilevel model was 9.0%. Because the multilevel model was fitted with all biomarkers as level-2 units, we did not perform cross-validation which may change the number of level-2 units retained for fitting the multilevel model.

For the random forest model, we included all the available predictors first. From the overall dataset, we obtained the feature importance for each of the predictors, ranked from the most to least relative contribution to decreasing the outcome variance at the tree splits where

the corresponding feature was featured. Among the predictors included, the number of affiliations from 3 to 5 year ago had the highest feature importance. Following the affiliation numbers, variance of SJR and growth rate of publication numbers from various years in the past, as well as the median SJR from 4 years ago, had relatively high feature importance. Further predictors along the ladder of importance appeared to be somewhat distanced from the previous predictors. To prevent the overfitting of the model to data, we conducted a 100-time cross-validation, essentially splitting the dataset randomly without replacement into a 75% training set and a 25% testing set for 100 times. During each split, the random forest model was fitted with the training set, while we recorded the median percentage deviation statistic obtained with the trained model applied on the corresponding testing set. The cross validation median percentage deviation for the random forest model ranged from 6.5% to 7.1%, with a mean of 6.8% and a median of 6.8%.

Table 2. Predictor feature importance from the random forest model

Predictor	Importance
Affiliation Number (3-year-ago)	5.91
Affiliation Number (5-year-ago)	4.22
Affiliation Number (4-year-ago)	4.16
Variance of SJR (3-year-ago)	2.70
Variance of SJR (5-year-ago)	2.43
Variance of SJR (4-year-ago)	2.26
Growth Rate of Adjusted Publication Number (4-year-ago)	2.13
Growth Rate of Adjusted Publication Number (5-year-ago)	1.94
Growth Rate of Adjusted Publication Number (3-year-ago)	1.74
Median SJR (4-year-ago)	1.66

Median SJR (3-year-ago)	1.33
Adjusted NIH Total Funding Amount (3-year-ago)	1.32
Median SJR (5-year-ago)	1.31
Adjusted NIH Project Number (3-year-ago)	1.27
Adjusted NIH Total Funding Amount (4-year-ago)	1.19
Adjusted NIH Project Number (4-year-ago)	1.03
Adjusted NIH Project Number (5-year-ago)	1.01
Adjusted NIH Median Funding Amount (4-year-ago)	0.98
Adjusted NIH Total Funding Amount (5-year-ago)	0.96
Adjusted NIH Median Funding Amount (3-year-ago)	0.92
Adjusted NIH Median Funding Amount (5-year-ago)	0.88
Biquadratic Year Number	0.83
Quadratic Year Number	0.82
Cubic Year Number	0.78
Year Number	0.76

To make future predictions on biomarker level, we refitted the random forest model with the full dataset in order to make future 3-years predictions for all the included biomarkers. The same set of predictions was conducted with the polynomial multilevel model, albeit allowing the predictions for future 5-years. The predictions for the popularity of the biomarkers are shown in the interactive plot previously discussed. In this interactive plot, viewers can visualize both the historical publication trend and future popularity prediction for either single biomarkers or multiple biomarkers comparatively. Viewers can also compare the predictive results from the two modeling frameworks we adopted.

Discussion

Using publicly available information for 90 distinct biomarkers between 2000 and 2020, we sought to document and understand the general trend of biomarker adoption in developmental science publications, as well as to explore the potential predictors behind their growth. We used a metascience perspective that was content agnostic, allowing us to identify common drivers to the popularity of biomarkers without regard to the substantive research topic. Most biomarkers followed the "learning curve" model, and though the rate of the growth differed by biomarker, there was a pattern of a rapid initial rate followed by slower, more consistent growth. In addition to documenting changes across time, we developed two models to predict the publication popularity for the biomarkers, finding that we were able to make relatively precise predictions for future popularity of the included biomarkers given historic publication and funding data. The two models were constructed from different modeling frameworks to present a diverse set of predictions and is less likely to be vulnerable to specific assumption violations. Both the explanation and the prediction components of the models offer important utilities. The explanations based on the interpretation of the models reveal what factors contributed to growth in biomarker adoption patterns. The predictions, on the other hand, offer insight into the trajectory of any given biomarker, which may be of interest for planning new projects or funding decisions.

In general, the patterns of publications most closely resembled a learning curve theory of growth in the field. The combination of the direction and magnitude of the three significant time related predictors, together resemble a general popularity growth curve that accelerates in the beginning, slows down in the middle, and is followed by a smaller accelerated growth after the slowdown. The consistency with which this general trend was found is impressive. However, it is possible that more transitory biomarkers saw so few publications as to never gain a foothold long enough to arrive on our list of biomarkers. Nevertheless, it does appear that developmental science incorporates biomarkers into research products in a remarkably similar way. In fact, only one biomarker, *5-HTT* or *SERT*, had a publication pattern that did not follow a learning

curve model. The majority of the papers under this heading are molecular genetic measures of the length polymorphism of the *5-HTT* gene (Caspi et al., 2003; Caspi, Hariri, Holmes, Uher, & Moffitt, 2010; Culverhouse et al., 2018; Karg, Burmeister, Shedden, & Sen, 2011; Munafò, Brown, & Hariri, 2008; Risch et al., 2009). The literature on candidate genes, and 5HTTLPR in particular, has been particularly active and contentious over a relatively short period of time (Alexander, 2019). After a very rapid rise, publications on *5-HTT* have seen a rapid decline recently suggesting a growth more akin to the infatuation cycle or potentially a hype curve (though more time would be needed to determine which pattern it most resembles).

Although most of the 90 biomarkers followed a similar trend in growth over the past 20 years, another key goal of this paper was to explain variations in that growth by using publicly accessible data on publications, funding, and networks from prior years. The purpose was in part to be able to describe some of the driving forces of the growth in these biomarkers. For example, the significant positive effect of the number of NIH projects funded fit expectations that a wide range of supported projects lead to future increases in publication output. However, the amount of funding was not a significant predictor over the number of projects, suggesting a weaker relation with future publication popularity.

Interestingly, we observed a significant negative effect of affiliation number from 5 years ago on publication popularity. This is contrary to expectations that topics that have already received widespread interest are in the position to gain even more popularity. Our results suggest that a rapid increase of participating research institutes is related to a decrease in future research popularity, lending support to the idea that a sudden influx of interest may precede a deceleration in the growth of research output. It is possible that as a biomarker becomes accessible for research by more institutions it, by definition, becomes less unique and innovative lowering its value for publication. Also, if more researchers can use the biomarker, it might also be more mature in its lifecycle. Scientists know its limitations and proper use, thereby reducing the number of publications with misuses of the biomarker. The feature importance

metric of the random forest model also demonstrated the importance of affiliation number as a driving force behind publication popularity. In this model, affiliation number had the highest contribution in explaining variance in publication popularity.

Furthermore, in terms of the random forest model, the feature importance metric also demonstrated that the variance of SJR index, and publication growth rate had the highest contributions to publication popularity. Unlike with the multilevel model, with the random forest model, current methods do not allow for us to elucidate the direction of relation between popularity and predictors. Thus, while we know that variance of SJR index and publication growth rate play an important role in publication popularity, we do not know whether a more or less variable SJR index or a lower or higher publication growth rate is associated with increased popularity. Based on our cross-validation process, the random forest model offered a higher prediction accuracy over the multilevel model, thus allowing for more confidence in the use of the prediction result for research activity and funding decision making. The random forest model, which depends on highly nonlinear relation, and the multilevel model, which uses largely linear relation between predictors and the outcome, draw information from distinctive sets of predictors. Because the two models rely on distinctive aspects of research to predict future publication popularity, users of our work have the freedom to pick the model that best captures their interest. For example, should a specific funding agency be particularly mindful of the variability in the research publication quality of a biomarker rather than how long the biomarker been published, the random forest predictions should be consulted more than the multilevel model.

In their current form, our models can inform funding agencies and researchers of the adoption trends for a wide range of biomarkers that covers almost all that have been associated with developmental research. Our users can quickly gather insight into the publication trends and predictions for single or multiple biomarkers. Furthermore, given the large number of biomarkers included in our modeling approach, a biomarker not included in our analysis, or

essentially any searchable scientific term, can have its future popularity predicted through feeding the models with the raw historic publication data.

Our data driven approach, while allowing us to model the publication popularity behavior with publicly available records, has limitations. Publication bias makes it more likely that whether studies translate to publications are based in part on whether the paper presents statistically significant results. Such biases are likely to mask the true use of each biomarker in research. Because we currently could not correct for publication biases in data, our predictive models could only rely on the assumption that all topics have similar biases in publication activities. With initiatives such as Open Science Framework, we project there will be metrics available that measure the degree of publication biases, thus allowing us to adjust the publication interest to reflect the true research interest popularity. Another limitation for our modeling approach involves the nature of our chosen predictors which are naturally correlated with high collinearity. Although we have taken care with the multilevel model to retain only a few key variables with a model comparison approach, it was nevertheless an incomplete approach that could result in violations of assumptions in the multilevel model. On the other hand, the random forest models should not be affected by the correlated predictors with the random selection of predictors at decision points. Taken together, future efforts should seek to address these weak points in both the source of data and the modeling approach.

Researchers who study infants, children, and adolescence have increasingly answered calls to be interdisciplinary (Duncan, 1991; Michel & Moore, 1995) and examine functioning across multiple levels (Falk et al., 2013; Insel, 2014; Pollak, 2015). Such approaches have already yielded important new knowledge related to normative development, the effects of adversity, and predictors of long-term life outcomes (e.g., morbidity and mortality). However, there are important considerations regarding biomarker adoption, particularly when team members have insufficient training to collect, assay, and interpret findings, especially in the context of small, cross-sectional samples and unspecified theoretical models. We hope to

encourage best practices within the developmental science community through the explicit encouragement of collaboration with those who have expertise and training in the biomarkers of interest, pre-registration and data sharing (in concert with the open science framework; Foster & Deardorff, 2017; Open Science Collaboration, 2015), and selection of biomarkers informed by theory. More broadly, research networks focused on improving the use of specific popular biomarkers, such as the Telomere Research Network and genetic and epigenetic consortia, bring together stakeholders to development and disseminate the best practices for telomere measurement. Taken together, our findings indicate that integration of biological and research on human development is likely to continue to grow – and it is encouraging that primarily patterns of learning, rather than hype curves, are best supported by patterns of adoption in the past 20 years.

References

Adler, P. S., & Clark, K. B. (1991). Behind the Learning Curve: A Sketch of the Learning Process on JSTOR. *Management Science*, 267–281.

Alexander, S. (2019, May). 5-HTTLPR: A Pointed Review. doi: 10.1002/JTS.21855

Baloff, N. (1966). The Learning Curve--Some Controversial Issues. *The Journal of Industrial Economics*, *14*(3), 275. doi: 10.2307/2097555

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models Usinglme4. *Journal of Statistical Software*, *67*(1). doi: 10.18637/jss.v067.i01

Berlyne, D. E. (1970). Novelty, complexity, and hedonic value. *Perception & Psychophysics 1970 8:5*, *8*(5), 279–286. doi: 10.3758/BF03212593

Bernard, H. R. (2012). The science in social science. *Proceedings of the National Academy of Sciences*, *109*(51), 20796–20799. doi: 10.1073/PNAS.1218054109

Bode, C., Herzog, C., Hook, D., & McGrath, R. (2018). *A Guide to the Dimensions Data Approach: A collaborative approach to creating a modern infrastructure for data describing research: Where we are and where we want to take it* .

Caspi, A., Hariri, A. R., Holmes, A., Uher, R., & Moffitt, T. E. (2010). Genetic Sensitivity to the Environment: The Case of the Serotonin Transporter Gene and Its Implications for Studying Complex Diseases and Traits. *The American Journal of Psychiatry*, *167*(5), 509. doi: 10.1176/APPI.AJP.2010.09101452

Caspi, A., Sugden, K., Moffitt, T., Taylor, A., Craig, I., Harrington, H., … Poulton, R. (2003). Influence of life stress on depression: moderation by a polymorphism in the 5-HTT gene. *Science (New York, N.Y.)*, *301*(5631), 386–389. doi: 10.1126/SCIENCE.1083968

Culverhouse, R., Saccone, N., Horton, A., Ma, Y., Anstey, K., Banaschewski, T., … Bierut, L. (2018). Collaborative meta-analysis finds no evidence of a strong interaction between stress and 5-HTTLPR genotype contributing to the development of depression. *Molecular Psychiatry*, *23*(1), 133–142. doi: 10.1038/MP.2017.44

Davis-Kean, P., & Ellis, A. (2019). An overview of issues in infant and developmental research for the creation of robust and replicable science. *Infant Behavior & Development*, *57*, 101339. doi: 10.1016/J.INFBEH.2019.101339

Dedehayir, O., & Steinert, M. (2016). The hype cycle model: A review and future directions. *Technological Forecasting and Social Change*, *108*, 28–41. doi: 10.1016/J.TECHFORE.2016.04.005

Duncan, G. J. (1991). Made in Heaven: Secondary Data Analysis and Interdisciplinary Collaborators. *Developmental Psychology*, *27*(6), 949–951. doi: 10.1037/0012-1649.27.6.949

Falk, E. B., Hyde, L. W., Mitchell, C., Faul, J., Gonzalez, R., Heitzeg, M. M., … Schulenberg, J. (2013). What is a representative brain? Neuroscience meets population science. *Proceedings of the National Academy of Sciences*, *110*(44), 17615–17622. doi: 10.1073/PNAS.1310134110

Fenn, J. (1995). When to Leap on the Hype Cycle. *Gartner*.

Fenn, J., & Linden, A. (2003). Understanding Gartner's Hype Cycles. *Gartner*.

Fenn, J., & Raskino, M. (2008). *Mastering the hype cycle: how to choose the right innovation at the right time*. Harvard Business Press.

Foster, E. D., & Deardorff, A. (2017). Open Science Framework (OSF). *Journal of the Medical Library Association*, *105*(2), 203–206. doi: 10.5195/JMLA.2017.88

Insel, T. (2014). The NIMH Research Domain Criteria (RDoC) Project: precision medicine for psychiatry. *The American Journal of Psychiatry*, *171*(4), 395–397. doi: 10.1176/APPI.AJP.2014.14020138

Justice, J., Ferrucci, L., Newman, A., Aroda, V., Bahnson, J., Divers, J., … Kuchel, G. (2018). A framework for selection of blood-based biomarkers for geroscience-guided clinical trials: Report from the TAME Biomarkers Workgroup. *Geroscience*, *40*(5–6), 419–436. doi: 10.1007/S11357-018-0042-Y

Karg, K., Burmeister, M., Shedden, K., & Sen, S. (2011). The serotonin transporter promoter variant (5-HTTLPR), stress, and depression meta-analysis revisited: evidence of genetic moderation. *Archives of General Psychiatry*, *68*(5), 444–454. doi: 10.1001/ARCHGENPSYCHIATRY.2010.189

Martin, A. R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B. M., & Daly, M. J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature Genetics*, *51*, 584–591. doi: 10.1038/s41588-019-0379-x

Meloni, M. (2014). How Biology Became Social, and What it Means for Social Theory. *The Sociological Review*, *62*(3), 593–614. doi: 10.1111/1467-954X.12151

Meloni, M. (2016). *Political Biology: Science and Social Values in Human Heredity from Eugenics to Epigenetics: https://doi.org/10.1177/0094306117744805bb*. New York, NY: Palgrave Macmillan. doi: 10.1177/0094306117744805BB

Michel, G. F., & Moore, C. L. (1995). *Developmental Psychobiology: An interdisciplinary science*. Cambridge, MA: MIT Press.

Mitchell, C, Schneper, L., & Notterman, D. (2016). DNA methylation, early life environment, and health outcomes. *Pediatric Research*, *79*(1–2), 212–219. doi: 10.1038/PR.2015.193

Mitchell, Colter, McLanahan, S., Brooks-Gunn, J., Garfinkel, I., Hobcraft, J., & Notterman, D. (2013). Genetic Differential Sensitivity to Social Environments: Implications for Research. *American Journal of Public Health*, *103*(Suppl 1), S102. doi: 10.2105/AJPH.2013.301382

Moore, S., & Kobor, M. (2020). Variability in DNA methylation at the serotonin transporter gene promoter: epigenetic mechanism or cell-type artifact? *Molecular Psychiatry*, *25*(9), 1906–1909. doi: 10.1038/S41380-018-0121-6

Munafò, M., Brown, S., & Hariri, A. R. (2008). Serotonin transporter (5-HTTLPR) genotype and amygdala activation: a meta-analysis. *Biological Psychiatry*, *63*(9), 852–857. doi: 10.1016/J.BIOPSYCH.2007.08.016

O'Leary, D. E. (2008). Gartner's hype cycle and information system research issues.

International Journal of Accounting Information Systems, *9*(4), 240–252. doi: 10.1016/J.ACCINF.2008.09.001

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251). doi: 10.1126/SCIENCE.AAC4716

Pollak, S. D. (2015). Multilevel developmental approaches to understanding the effects of child maltreatment: Recent advances and future challenges. *Development and Psychopathology*, *27*(4 Pt 2), 1397. doi: 10.1017/S0954579415000826

Risch, N., Herrell, R., Lehner, T., Liang, K.-Y., Eaves, L., Hoh, J., … Merikangas, K. R. (2009). Interaction between the serotonin transporter gene (5-HTTLPR), stressful life events, and risk of depression: a meta-analysis. *JAMA*, *301*, 2462–2471. doi: 10.1001/jama.2009.878

Ritchie, S. (2020). *Science Fictions: How Fraud, Bias, Negligence, and Hype Undermine the Search for Truth*. New York, NY: Metropolitan Books.

Singh, I., & Rose, N. (2009). Biomarkers in psychiatry. *Nature*, *460*(9), 202–207. doi: 10.1038/460202a

Thornton, A. (2005). *Reading History Sideways: The Fallacy and Enduring Impact of the Developmental Paradigm on Family Life*. Chicago, IL: The University of Chicago Press.