

## Appendix

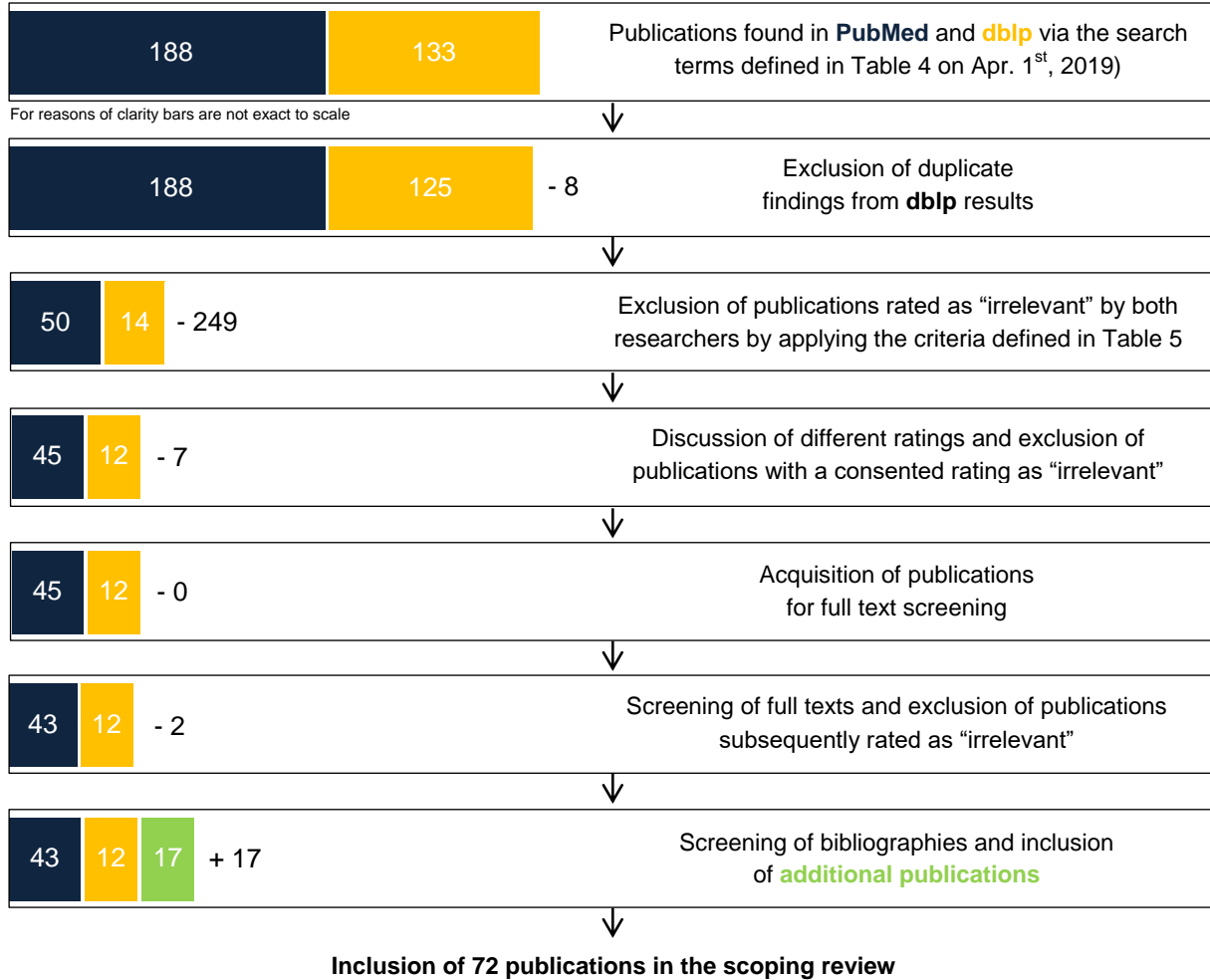


Figure 4: Graphical representation of the scoping review process

## Search terms and data

Database	Data type	Search term	Language	Search Fields	Limiters	Search results (Apr. 1 <sup>st</sup> , 2019)	Proportion
Pub Med	Electronic health record data	(big data analy* [Title/Abstract] OR advanced analy* [Title/Abstract] OR predictive analy* [Title/Abstract] OR machine learning [Title/Abstract] AND (EHR [Title/Abstract] OR EMR [Title/Abstract] OR electronic health record* [Title/Abstract] OR electronic medical record* [Title/Abstract]))	English	Title/ Abstract	Journals	467	5.45%
Pub Med	Transactional claims data	((big data analy* [Title/Abstract] OR advanced analy* [Title/Abstract] OR predictive analy* [Title/Abstract] OR machine learning [Title/Abstract]) AND (claim* [Title/Abstract] OR administrative data [Title/Abstract] OR transactions data [Title/Abstract] OR billing data [Title/Abstract] OR health business data [Title/Abstract]))	English	Title/ Abstract	Journals	178	2.08%
Pub Med	Data from clinical registries	((big data analy* [Title/Abstract] OR advanced analy* [Title/Abstract] OR predictive analy* [Title/Abstract] OR machine learning [Title/Abstract]) AND (registry [Title/Abstract] OR registries [Title/Abstract]))	English	Title/ Abstract	Journals	99	1.16%
Pub Med	Smart sensor and smart device data	((big data analy* [Title/Abstract] OR advanced analy* [Title/Abstract] OR predictive analy* [Title/Abstract] OR machine learning [Title/Abstract]) AND (device* [Title/Abstract] OR sensor* [Title/Abstract] OR smartphone* [Title/Abstract] OR wearable* [Title/Abstract]))	English	Title/ Abstract	Journals	1,370	15.99%
Pub Med	Internet usage and social media data	((big data analy* [Title/Abstract] OR advanced analy* [Title/Abstract] OR predictive analy* [Title/Abstract] OR machine learning [Title/Abstract])	English	Title/ Abstract	Journals	223	2.60%

		AND (social media [Title/Abstract] OR web search [Title/Abstract] OR health portal [Title/Abstract] OR health information [Title/Abstract]))					
Pub Med	Data from clinical trials	((big data analy* [Title/Abstract] OR advanced analy* [Title/Abstract] OR predictive analy* [Title/Abstract] OR machine learning [Title/Abstract]) AND (clinical trial [Title/Abstract] OR RCT [Title/Abstract] OR controlled trial [Title/Abstract]))	English	Title/ Abstract	Journals	137	1.60%
Pub Med	Imaging data	((big data analy* [Title/Abstract] OR advanced analy* [Title/Abstract] OR predictive analy* [Title/Abstract] OR machine learning [Title/Abstract]) AND (*imaging [Title/Abstract] OR images [Title/Abstract]))	English	Title/ Abstract	Journals	2,644	30.86%
Pub Med	Biomarker data	((big data analy* [Title/Abstract] OR advanced analy* [Title/Abstract] OR predictive analy* [Title/Abstract] OR machine learning [Title/Abstract]) AND (omics [Title/Abstract] OR gene [Title/Abstract] OR genetic [Title/Abstract] OR DNA [Title/Abstract] OR Biomarker [Title/Abstract]))	English	Title/ Abstract	Journals	3,370	39.34%
Pub Med	Health- related systems data	((big data analy* [Title/Abstract] OR advanced analy* [Title/Abstract] OR predictive analy* [Title/Abstract] OR machine learning [Title/Abstract]) AND (socio-economic data [Title/Abstract] OR census [Title/Abstract] OR location data [Title/Abstract] OR environmental data [Title/Abstract]))	English	Title/ Abstract	Journals	52	0.61%
Pub Med	Patient- reported outcomes and survey data	((big data analy* [Title/Abstract] OR advanced analy* [Title/Abstract] OR predictive analy* [Title/Abstract] OR machine learning [Title/Abstract]) AND (patient reported outcome [Title/Abstract]	English	Title/ Abstract	Journals	27	0.32%

		OR PROM [Title/Abstract] OR survey data [Title/Abstract])					
--	--	--	--	--	--	--	--

Table 4: Search terms for the rapid review on big data types in healthcare

Database	Data type	Search term	Language	Search Fields	Limiters	Search results (Apr. 1 <sup>st</sup> , 2019)	Proportion
PubMed	Logistic regression	((big data analy* [Title/Abstract] OR advanced analy* [Title/Abstract] OR predictive analy* [Title/Abstract] OR machine learning [Title/Abstract]) AND ("logistic regression" [Title/Abstract]))	English	Title/ Abstract	Journals	1,026	12.03%
PubMed	Linear regression	((big data analy* [Title/Abstract] OR advanced analy* [Title/Abstract] OR predictive analy* [Title/Abstract] OR machine learning [Title/Abstract]) AND ("linear regression" [Title/Abstract]))	English	Title/ Abstract	Journals	312	3.66%
PubMed	Neural network	((big data analy* [Title/Abstract] OR advanced analy* [Title/Abstract] OR predictive analy* [Title/Abstract] OR machine learning [Title/Abstract]) AND ("neural network" [Title/Abstract]))	English	Title/ Abstract	Journals	1,742	20.43%
PubMed	Bayesian network	((big data analy* [Title/Abstract] OR advanced analy* [Title/Abstract] OR predictive analy* [Title/Abstract] OR machine learning [Title/Abstract]) AND ("Bayesian network" [Title/Abstract]))	English	Title/ Abstract	Journals	123	1.44%
PubMed	Decision tree	((big data analy* [Title/Abstract] OR advanced analy* [Title/Abstract] OR predictive analy* [Title/Abstract] OR machine learning [Title/Abstract]) AND ("decision tree" [Title/Abstract]))	English	Title/ Abstract	Journals	569	6.67%
PubMed	k-nearest neighbour	((big data analy* [Title/Abstract] OR advanced analy* [Title/Abstract] OR predictive analy* [Title/Abstract] OR machine learning [Title/Abstract]) AND ("k-nearest" [Title/Abstract]))	English	Title/ Abstract	Journals	517	6.06%
PubMed	Random forest	((big data analy* [Title/Abstract] OR advanced analy* [Title/Abstract])	English	Title/ Abstract	Journals	1,664	19.51%

		OR predictive analy* [Title/Abstract] OR machine learning [Title/Abstract] AND ("random forest" [Title/Abstract]))					
PubMed	k-means clustering	((big data analy* [Title/Abstract] OR advanced analy* [Title/Abstract] OR predictive analy* [Title/Abstract] OR machine learning [Title/Abstract] AND ("k-means" [Title/Abstract]))	English	Title/ Abstract	Journals	163	1.91%
PubMed	Support vector machine	((big data analy* [Title/Abstract] OR advanced analy* [Title/Abstract] OR predictive analy* [Title/Abstract] OR machine learning [Title/Abstract] AND ("support vector machine" [Title/Abstract]))	English	Title/ Abstract	Journals	2,327	27.29%
PubMed	Regularized regression	((big data analy* [Title/Abstract] OR advanced analy* [Title/Abstract] OR predictive analy* [Title/Abstract] OR machine learning [Title/Abstract] AND ("regularized regression" [Title/Abstract]))	English	Title/ Abstract	Journals	22	0.26%
PubMed	Multiple regression	((big data analy* [Title/Abstract] OR advanced analy* [Title/Abstract] OR predictive analy* [Title/Abstract] OR machine learning [Title/Abstract] AND ("multiple regression" [Title/Abstract]))	English	Title/ Abstract	Journals	43	0.50%
PubMed	Proportional hazard model	((big data analy* [Title/Abstract] OR advanced analy* [Title/Abstract] OR predictive analy* [Title/Abstract] OR machine learning [Title/Abstract] AND ("proportional hazard" [Title/Abstract]))	English	Title/ Abstract	Journals	19	0,22%

Table 5: Search terms for the rapid review on analytical methods for BDA in healthcare

Database	Search term	Language	Search Fields	Limiters	Search results (Apr. 1 <sup>st</sup> , 2019)
PubMed	( big data analy*[Title/Abstract] OR advanced analy*[Title/Abstract] OR predictive analy*[Title/Abstract] OR (machine learning [Title/Abstract]	English	Title/ Abstract	Case report, guideline, observational study, review,	188

	AND observational [Title/Abstract]) ) NOT (genetic* [Title/Abstract] OR chemic* [Title/Abstract] OR biologic* [Title/Abstract] OR *omic* [Title/Abstract] OR molecu* [title/abstract] OR clinical trial [Title/Abstract] OR natural language processing [Title/Abstract] )			systematic review, technical report	
dblp	Big data analy* health Advanced analy* health Predictive analy* health Visual analy* health Machine learning health	English	All	Journal article	133

Table 6: Final search terms for the scoping review

Criteria	Included	Excluded
Dates	≥ 2013	< 2013
Language	English	All other languages
Article type	(Systematic) reviews, case studies, discussion papers, guidelines, method papers, technical reports, research papers with BDA application	Books, qualitative research, clinical trials, editorials, conference supplements, research papers without BDA application, etc.
Content	<ul style="list-style-type: none"> <li>- big data platform</li> <li>- big data technologies</li> <li>- health data types/sources</li> <li>- analytical models for big data</li> <li>- BDA applications/use cases</li> <li>- BDA challenges/ potentials</li> </ul>	<ul style="list-style-type: none"> <li>- Natural language processing</li> <li>- big data only in the outlook</li> <li>- genetic/molecular/etc. research</li> <li>- no big data and no advanced analytics</li> <li>- too specific, not directly human health-related context, e.g. BDA for animal research, crop science or air pollution</li> </ul>
Filter	machine learning for observational research	machine learning in general as most often used in genomic studies

Table 7: Inclusion and exclusion criteria for the scoping review

Strategy	Strategic Intervention	References mentioning BDA as supporting tool	Number and % of
----------	------------------------	--	-----------------

domain			References (n= 72)	
Empowering and engaging people	Personalized care plans	[8–10, 14, 15, 24, 26, 28, 40, 42–44, 46, 47, 51, 52, 56, 59, 72, 74, 75, 87–89, 93–95, 98, 104, 116, 124]	31	43.1%
	Self-management activities	[22, 43, 44, 50, 74]	5	6.9%
	Shared decision making	[40, 51, 105, 125]	4	5.6%
	Health education	[15, 46, 89]	3	4.2%
	Access to personal health records	[71, 105]	2	2.8%
	Peer support	[43]	1	1.4%
	Patient satisfaction surveys	[74]	1	1.4%
Strengthening governance and accountability	Performance evaluation	[10, 26, 27, 45, 46, 54, 74, 79, 91, 93, 98, 100, 101, 103, 104]	15	20.8%
	Performance-based contracting	[26, 27, 45, 79, 93, 98, 101, 104]	8	11.1%
	Decentralization	[9, 43, 74, 92, 93, 100, 124, 125]	8	11.1%
	Patient-reported outcomes	[51]	1	1.4%
Reorienting the model of care	Clinical decision support	[7, 10, 14, 15, 24, 26, 27, 36, 40, 46, 48, 58, 59, 74, 76, 87, 90, 93, 95, 105, 114, 126, 127]	23	31.9%
	Population-based services	[8, 14, 15, 22, 24, 26, 27, 46, 52, 56, 59, 61, 80, 87, 91, 93, 94, 124, 128]	19	26.4%
	Surveillance and control systems	[8, 14, 43, 44, 46, 47, 54, 58, 74, 88, 91, 93, 125]	13	18.1%
	Mobile health technologies	[7, 15, 38, 40, 41, 71, 74, 89, 104, 108]	10	13.9%
	Health promotion and disease prevention	[22, 28, 51, 85, 89, 98, 115, 116, 129]	9	12.5%
	Home and nursing care	[14, 50, 60, 72, 89]	5	6.9%
Coordinating services	Care pathways	[22, 24, 61–63, 80, 124, 128]	8	11.1%
	Sharing of medical records	[10, 41, 71, 76, 77, 105]	6	8.3%
	Intersectoral partnerships	[10, 45, 50, 74, 116]	5	6.9%
	District-based healthcare delivery	[100]	1	1.4%
Creating an enabling environment	Resource allocation	[15, 37, 38, 40, 45, 51, 54, 79, 101, 105, 108]	11	15.3%
	System research	[9, 46, 50, 75, 93, 104]	6	8.3%

	Quality assurance	[14, 39, 113]	3	4.2%
	Workforce training	[74, 106]	2	2.8%

Table 8: Matrix table on the strategic interventions of the people-centred and integrated health services framework supported by big data analytics in this scoping review

Challenge domain	Challenge	References mentioning BDA challenges (n = 45 (63.3%) of 71)	Number and % of references (n = 72)	
Regulatory	Investment framework	[7, 27, 36, 58, 77, 85, 104]	7	9.7%
	Communication framework	[9, 36, 41, 103–105]	6	8.3%
	Intellectual property framework	[9, 14, 15, 36, 39, 41, 44, 72, 77, 88, 108]	11	15.3%
	Evaluation framework	[41, 46, 52, 72, 75, 76, 85, 104–106, 115, 116, 124]	13	18.1%
	Privacy framework	[14, 15, 27, 35–37, 41, 44, 46, 52, 54, 71, 75, 77, 85, 88, 89, 93, 105, 124]	20	27.8%
	Ethics framework	[7, 41, 88, 104, 108]	5	6.9%
Technological	Data infrastructure	[9, 15, 27, 36–38, 46, 59, 75, 77, 85, 89, 104, 108]	14	19.4%
	Data processing	[9, 14, 27, 28, 36, 38, 46, 59, 72, 75, 88, 104, 108, 124]	14	19.4%
	Data linkage	[14, 27, 28, 36–39, 44, 46, 54, 72, 77, 85, 88, 89, 104, 108, 116, 124]	19	26.4%
	Data quality	[7, 10, 14, 27, 37–39, 41, 44, 46, 54, 72, 75, 77, 85, 93, 104, 108, 125]	19	26.4%
	Data access & security	[9, 14, 15, 27, 37, 38, 44, 46, 52, 54, 71, 72, 75, 77, 88, 93, 108, 124]	18	25.0%
Methodological	High-dimensional analytics	[9, 44, 46]	3	4.2%
	Real-time analytics	[14, 36, 38, 44, 72, 75, 77, 85, 88, 89, 104, 105, 124, 127]	14	19.4%
	Modelling standards & bias	[7, 10, 14, 27, 36, 39, 44, 46, 54, 59, 72, 75, 77, 85, 87, 89, 90, 92, 93, 104, 108, 113–115, 124, 125]	26	36.1%
	Evidence-base	[7, 9, 27, 36, 39, 41, 44, 46, 48, 52, 54, 56, 58, 72, 75–77, 89, 93, 104, 115, 116, 124, 125]	24	33.3%
	Interpretation & usability	[7, 14, 27, 46, 48, 58, 59, 71, 72, 85, 90, 91, 93, 104, 115, 124, 125]	17	23.6%
Cultural	Teamwork culture	[9, 27, 36, 37, 46, 58, 59, 72, 75, 85, 89, 105]	12	16.7%
	Process redesign	[15, 27, 36, 40, 46, 72, 105]	7	9.7%
	Data sharing culture	[15, 36, 41, 44]	4	5.6%
	Data governance	[27, 72]	2	2.8%
	Learning & change culture	[9, 27, 37, 58, 72, 77, 85, 93, 105]	9	12.5%

Table 9: Main challenges of BDA discussed in the screened literature



<b>Country assignment of the first author</b>	<b>Number of publications in the review (n = 72)</b>
<i>USA</i>	43
<i>Canada</i>	3
<i>United Kingdom</i>	3
<i>Australia</i>	2
<i>China</i>	2
<i>Denmark</i>	2
<i>Germany</i>	2
<i>Saudi Arabia</i>	2
<i>Spain</i>	2
<i>Switzerland</i>	2
<i>Austria</i>	1
<i>Belgium</i>	1
<i>India</i>	1
<i>Italy</i>	1
<i>Malaysia</i>	1
<i>Poland</i>	1
<i>South Africa</i>	1
<i>Sweden</i>	1
<i>Turkey</i>	1

Table 10: Publications included in the scoping review by country

<b>Study type</b>	<b>Number of publications in the review (n = 72)</b>
<i>Review</i>	24
<i>Case report</i>	17
<i>Quantitative analysis</i>	13
<i>Technical report</i>	12
<i>Guideline</i>	5
<i>Survey</i>	1

Table 11: Publications included in the scoping review by study type

<b>Research setting</b>	<b>Number of publications in the review (n = 72)</b>
<i>Research</i>	34
<i>Hospital</i>	15
<i>Population health management</i>	14
<i>Health insurance</i>	5
<i>Pharmaceuticals</i>	4

<i>Public health</i>	2
<i>Community care</i>	1

*Table 12: Publications included in the scoping review by setting*

## References

124. Belle A, Thiagarajan R, Soroushmehr SMR, et al. Big data analytics in healthcare. *BioMed Research International*. 2015; 1–16. DOI: <https://doi.org/10.1155/2015/370194>
125. Bragazzi NL, Alicino C, Trucchi C, et al. Global reaction to the recent outbreaks of Zika virus: insights from a big data analysis. *PLOS ONE*. 2017; 12(9): e0185263. DOI: <https://doi.org/10.1371/journal.pone.0185263>
126. Kaewprag P, Newton C, Vermillion B, et al. Predictive models for pressure ulcers from intensive care unit electronic health records using Bayesian networks. *BMC Med Inform Decis Mak*; 2017. DOI: <https://doi.org/10.1186/s12911-017-0471-z>
127. Taylor RA, Pare JR, Venkatesh AK, et al. Prediction of in-hospital mortality in emergency department patients with sepsis: a local big data-driven, machine learning approach. *Academic Emergency Medicine*. 2016; 23(3): 269–278. DOI: <https://doi.org/10.1111/acem.12876>
128. Makar M, Ghassemi M, Cutler DM, Obermeyer Z. Short-term mortality prediction for elderly patients using medicare claims data. *International Journal of Machine Learning and Computing*. 2015; 5(3): 192–197. DOI: <https://doi.org/10.7763/IJMLC.2015.V5.506>
129. Bihorac A, Ozrazgat-Baslanti T, Ebadi A, et al. MySurgeryRisk: development and validation of a machine-learning risk algorithm for major complications and death after surgery. *Annals of Surgery*. 2018; 1. DOI: <https://doi.org/10.1097/SLA.0000000000002706>