보건·복지 Issue & Focus

Korea Institute for Health and Social Affairs

ISSN 2092-7117
제168호(2012-49) 발행일 : 2012. 12. 14

KIHASA 한국보건사회연구원 Korea Institute for Health and Social Affairs

빅 데이터를 활용한 자살요인 다변량 분석: Google 검색트렌드 적용

우리나라의 보건복지 분야에서는 이미 수많은 빅 데이터가 정부 및 공공기관에서 관리되고 있으나 정보 접근의 어려움으로 새로운 가치를 창출할 수 있는 빅 데이터 활용에는 미흡함

빅 데이터를 통한 자살요인 분석은 구글의 검색트랜드(http://www.google.co.kr/trends/)를 이용하여 한국의 자살 검색량의 결정요인에 대한 다변량 분석을 실시한 결과, 한국의 자살률은 구글의 자살 검색량과 비슷한 추세를 보이고 있고, 연도별 자살률은 자살 검색량을 유의하게 높이는 것으로 나타남

본 고의 스트레스 검색량이 자살 검색량에 직접적인 영향을 주는 것으로 나타나 민간 검색포털이나 SNS에서 관련 키워드의 검색이나 Buzz 발생 시 징후가 예측되면 이용자에게 가장 적합한 스트레스 관리 프로그램을 팝업창이나 문자 메시지를 통하여 제공함으로써 자살충동을 예방하는 데 기여할 수 있을 것임

송태민 연구위원

1. 빅 데이터 활용의 필요성

■ 빅 데이터란?

○스마트기기, 센서 등의 급속한 보급과 모바일 인터넷과 SNS의 확산으로 데이터량이 기하급수적으로 증가하여 데이터가 경제적 자산이 될 수 있는 빅 데이터 시대가 도래

○2011년 전 세계 데이터에서 생성될 디지털 정보량이 1.8ZB(제타바이트)에 달하는 '제타바이트 시대'로 진입[1)]

– 1ZB는 1조GB(기가바이트)에 해당하는 양으로 미의회도서관 저장정보(235TB(테라바이트), '11. 4월 현재)의 약 4백만배에 해당

1) 윤형중(2012). 이제는 빅 데이터 시대. e비즈북스.

○빅 데이터는 엄청나게 많은 데이터로 양적인 의미를 벗어나 데이터의 분석과 활용을 포괄하는 개념

– 빅 데이터는 기존의 데이터베이스 시스템으로는 수집 · 저장 · 관리 · 분석하기 힘든 대량의 정형 또는 비정형의 데이터로부터 새로운 가치를 창출하는 기술로 정의

■ 빅 데이터의 가치[2)]

○세계 각국의 정부와 기업들은 빅 데이터가 향후 국가와 기업의 성패를 가름할 새로운 경제적 가치의 원천이 될 것으로 기대하며 다양한 부문에서 빅 데이터의 적극적인 활용을 시도

○맥킨지(McKinsey)는 의료, 공공행정, 소매, 제조, 개인정보 등 다양한 부문에 빅 데이터의 적용이 가능할 것으로 예측

– 미국의 경우 최대 7천억 달러의 경제적 효과가 창출되고 EU는 연간 1,500억 유로 이상, 한국은 10.7조 원 이상 정부지출이 감소될 것으로 예측

○가트너(Gartner)는 빅 데이터가 미래 경쟁력을 좌우하는 21세기의 원유로 수많은 일자리 창출을 전망

– 2015년까지 전 세계적으로 440만개, 미국 내에서만 190만개의 빅 데이터 관련 일자리가 창출될 것으로 전망

■ 빅 데이터의 활용사례

○보건의료분야에서는 많은 국가의 공공부분에서 유전자와 생명연구자원의 공유를 통하여 질병의 예방(예측)과 치료, 환자관리 등에 활용하고 있으며 다국적 IT기업과 검색포털들은 기존에 저장된 빅 데이터를 분석하여 다양한 가치 정보를 제공

– 미국 국립보건원은 제조사와 사용자 간의 쌍방향 상호작용을 통해 사용자가 요구하는 다양한 약에 대한 정보를 제공하는 Pillbox서비스(pillbox.nlm.nih.gov)를 구축하여 연간 5,000만달러의 비용절감[3)]

– 구글은 사용자가 입력한 검색어의 로그를 분석하여 독감예보 서비스(www.google.org/flutrends)를 제공하여 전 세계 독감확산 현상을 실시간으로 제공

– 한국정보화진흥원 빅 데이터 국가전략포럼 분석팀에서는 2012년 1월부터 10월 18일까지 자살로 언급된 빅 데이터 자료를 뉴스, 블로그, 카페, SNS, 게시판 등에서 수집하여 청소년이 작성했다고 추정되는 69,886건의 자료를 분석하여 청소년들의 자살과 관련한 온라인 Buzz의 발생패턴에 따라 보다 체계적으로 대응할 수 있는 자살예방체계를 설계할 수 있다는 가능성을 보임[4)]

■ 자살요인 분석의 빅 데이터 활용

○우리나라의 보건복지 분야에서는 이미 수많은 빅 데이터가 정부 및 공공기관에서 관리되고 있으나 정보 접근의 어려움으로 새로운 가치를 창출할 수 있는 빅 데이터 활용에는 미흡함

– 공공부문의 빅 데이터는 정보의 제한적 서비스로 활용이 미흡한 반면, 민간기관의 검색포털이나 SNS에서 관리되고 있는 빅 데이터의 분석과 활용은 활발히 이루어지고 있음

○우리나라는 급격한 사회 · 경제적 변화속에 자살률이 2004년부터 OECD 국가 중 최고의 수준이며, 특히 청소년계층의 자살 문제가 사회적 이슈로 대두되면서 정부차원의 적극적인 대책이 시급한 실정임

2) '한국정보화진흥원(2012. 10. 29). 대한민국 사회현안과 빅 데이터 전략' 과 '국가정보화전략위원회(2011. 11). 지식정보 개방과 협력으로 스마트정부 구현' 에서 인용

3) 한국정보화진흥원(2012). 빅 데이터로 진화하는 세상(Big data 글로벌 선진 사례).

4) 한국정보화진흥원(2012. 10. 29). 대한민국 사회현안과 빅 데이터 전략.

○그동안 자살 연구는 국가 간 자살률 비교나 패널데이터의 분석을 통한 자살원인의 요인으로 정신과적 요인, 생물학 및 의학적 요인, 사회환경적 요인 등에 초점을 맞춘 연구가 진행됨

– 사회환경적 요인 중 유명연예인이나 또래 친구의 자살을 모방하는 '베르테르 효과' 도 자살의 위험을 높일 수 있는 원인이며, 스트레스가 우울을 유발하여 자살의 가능성을 높일 수 있다고 보고됨

○패널 데이터의 분석은 데이터 수집의 제한으로 인하여 개인과 집단의 다양한 자살원인에 대한 분석은 미흡한 실정임

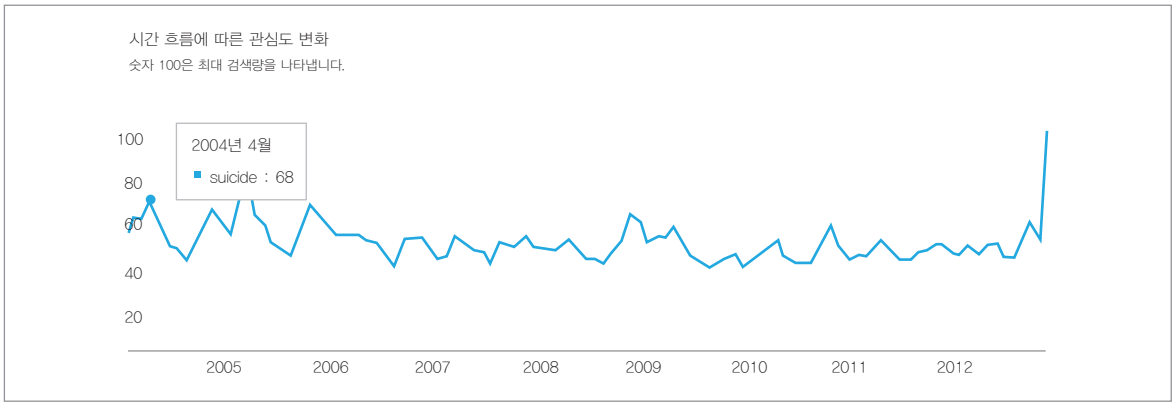
– 본 고에서는 구글의 검색트랜드를 활용하여 자살요인에 대한 다변량 분석을 적용해 봄으로써 보건복지분야 빅 데이터의 활용방안을 제시함

○빅 데이터를 통한 자살요인 분석은 구글의 검색트랜드(http://www.google.co.kr/trends/)를 이용하여 한국의 자살 검색량의 결정요인에 대한 다층(위계적 선형)모형 분석과 한국과 미국의 자살요인에 대한 집단 간 구조모형 분석을 실시함

– 구글의 검색트랜드는 전 세계의 사용자가 입력한 검색어 빅 데이터를 분석하여 사용자가 특정시간에 특정 지역에서 검색어를 입력한 검색량을 표준화하여 제공함

– 본 연구에 사용된 구글의 검색어로는 스트레스(stress), 음주(drinking), 운동(exercise), 자살(suicide)로 2004년 1월부터 2012년 10월까지의 검색량을 이용함[5)]

[그림 1] 구글 검색트랜드(http://www.google.co.kr/trends/)의 suicide 검색어 검색 결과(미국)



2. OECD 국가별 자살률과 빅 데이터 자살 검색량 비교

■ OECE 주요 국가의 자살률과 자살 검색량 비교

○2010년 OECD 회원국의 자살에 의한 평균 사망률은 인구 10만명당 12.8명이며, 우리나라는 33.5명으로 회원국 중 가장 높음(OECD Health Data 2012)

○주요 OECD 국가의 자살률과 자살 검색량을 비교한 결과 대부분 국가의 자살률과 자살 검색량은 안정적으로 지속되거나 감소하였으나, 한국의 자살률과 자살 검색량은 증가 추세를 보이는 것으로 나타남

5) 국내의 횡단적 연구와 종단적 연구를 통하여 스트레스는 우울과 자살에 밀접한 관련이 있으며, 스트레스와 자살 간에 건강생활실천요인(음주, 운동 등)이 매개하여 영향을 준다는 결과에 따라 본 고의 자살요인 분석에는 4개의 키워드를 선정함

〈표 1〉 OECD 주요 국가의 자살률과 자살 검색량 비교

국가	2005년		2006년		2007년		2008년		2009년		2010년	
	자살률[1]	검색량[2]	자살률[1]	검색량[2]	자살률[1]	검색량[2]	자살률[1]	검색량[2]	자살률[1]	검색량[2]	자살률[1]	검색량[2]
독일	11.4	42.8	10.7	41.2	10.2	40.3	10.3	41.8	10.3	41.3	10.8	41.0
미국	11.2	74.2	11.3	64.2	11.7	59.7	12.0	61.0	–	57.5	–	57.4
스웨덴	13.1	79.3	12.7	75.3	11.9	64.1	12.2	61.4	12.9	57.1	11.7	61.0
영국	6.7	87.0	6.7	77.0	6.3	63.8	6.9	67.0	6.8	56.8	6.7	53.4
이탈리아	–	59.5	5.6	72.1	5.7	57.3	5.8	48.3	5.9	40.3	–	34.3
일본	22.1	71.2	21.6	62.0	22.1	59.3	21.8	69.8	22.2	76.4	21.2	83.3
프랑스	17.1	63.4	16.5	52.9	15.8	52.5	16.1	53.6	16.2	49.9	–	49.3
핀란드	18.3	49.0	19.6	43.7	18.2	42.8	19.0	46.8	18.9	47.7	17.3	44.0
호주	10.3	79.1	10.4	69.6	10.8	60.2	10.8	54.5	10.5	49.9	10.6	51.0
한국	29.9	58.9	26.2	42.2	28.7	47.3	29.0	55.8	33.8	58.8	33.5	78.3

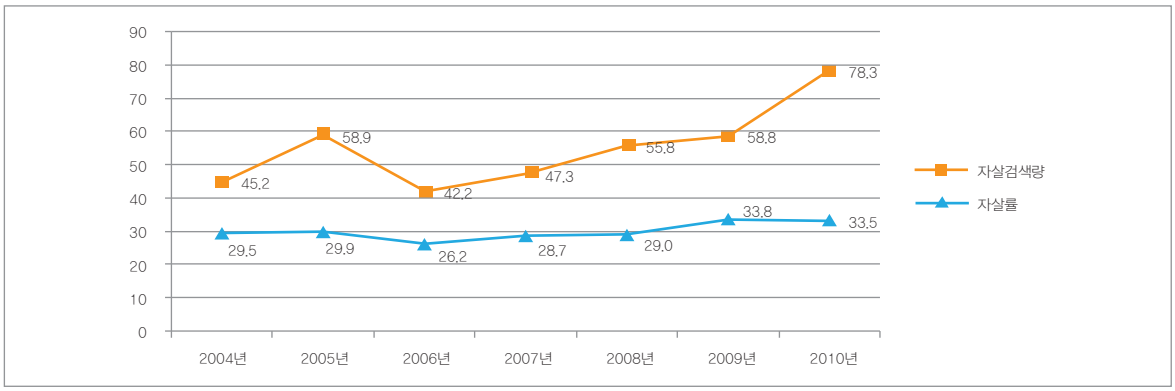
1) 자살률은 인구 100,000명당 자살 수를 나타냄(OECD Health Data 2012)
2) 자살 검색량은 google에서 실행된 총 검색수 대비 자살용어(suicide, 자살)에 대한 검색수로 특정 시간에 특정 지역에서 검색할 확률을 나타냄

■ 우리나라의 자살률과 자살 검색량 비교

○한국의 자살률은 2003년 28.1명에서 2004년 29.5명, 2005년 29.9명으로 증가하다가 2006년 감소하였다가 2007년 이후 증가추세를 보이고 있으며, 구글의 자살 검색량도 자살률과 비슷한 추세를 보이고 있음

– 특히, 2005년과 2008년 유명 연예인의 자살과 함께 자살률과 자살 검색량이 증가한 것은 모방자살에 대한 위험이 높은 것으로 나타남

[그림 2] 한국의 자살률과 자살 검색량 비교



○연도별 자살 검생량의 결정요인에 대한 다층(위계적 선형)모형 분석결과 연도별 자살률은 자살 검색량을 유의하게 높이는 것으로 나타남

– 연도별 자살 검색량의 통계적 차이를 분별하는 모형 1(기초모형)에서 고정효과(Fixed Effect)는 55.202로 유의하게(p〈.001) 나타났으며, 집단내 상관계수(ICC: Interclass Correlation)는 .332로 유의하게 나타나 (χ^2=41.909, P〈.001) 연도별 특성에 따라 자살 검색량이 33.2% 정도 영향을 받는 것으로 나타남

– 모형 2(무조건적 기울기 모형) 검증에서 월별 자살 검색량의 영향에 있어 스트레스 검색량이 .38 정도 자살 검색량을 유의하게(p〈.1) 높이는 것으로 나타났으며, 임의효과(Random Effect)도 유의미하게(p〈.1) 나타나 스트레스 검색량도 연도별 차이가 있는 것으로 나타남

– 모형 3(조건적 모형) 검증에서 연도별 자살률은 자살 검색량을 3.26정도 유의하게(P<.05) 높이는 것으로 나타남(mixed model: 자살검색량$_{ij}$=γ_{00} + γ_{01}*자살률$_j$+γ_{10}*스트레스검색량$_{ij}$+γ_{20}*음주검색량$_{ij}$+γ_{30}*운동검색량$_{ij}$+u_{0j}+u_{1j}*스트레스검색량$_{ij}$+u_{2j}*음주검색량$_{ij}$+u_{3j}*운동검색량$_{ij}$+r_{ij})

3. 자살요인에 대한 다중집단 구조모형 분석

■ 자살요인 다중집단 구조모형의 적합성[6)]

○자살요인의 다중집단 구조모형 분석은 연구모형의 적합성을 검증한 후, 집단 간 등가제약 과정을 거쳐 경로계수 간 유의미한 차이를 검증함

– 다중집단분석 연구모형의 적합도는 χ^2 (df, p)=2.061(2, .357), GFI=.995, NFI=0.991, TLI=0.998, RMSEA=0.012로 모든 적합도에서 적합한 것으로 나타남

■ 자살요인 다중집단 경로분석

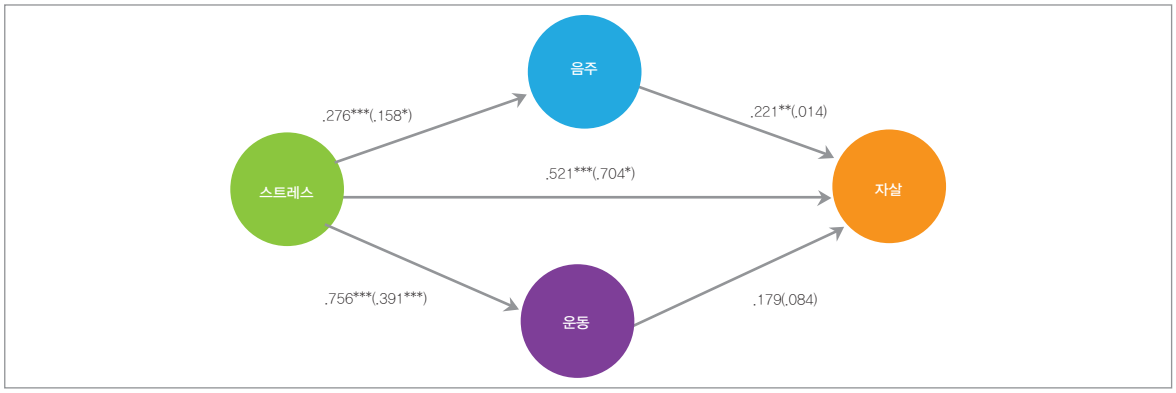
○한국과 미국 두 집단 모두 스트레스에서 운동, 음주, 그리고 자살로 가는 경로에 정적(+)으로 유의한 영향을 미치는 것으로 나타남

〈표 2〉 한국과 미국의 집단별 모수 추정치

경로	한국			미국		
	B(β)	C.R.	P	B(β)	C.R.	P
스트레스→음주	.186(.276)	2.941	***	.172(.158)	1.644	*
스트레스→운동	.777(.756)	11.834	***	.337(.391)	4.357	***
음주→자살	.186(.222)	2.538	**	.013(.014)	.211	.833
운동→자살	.099(.179)	1.395	.163	.099(.084)	1.179	.239
스트레스→자살	.142(.251)	1.922	*	.717(.704)	9.820	***

B: 비표준화회귀계수 β : 표준화회귀계수
*** p<0.01, ** p<0.05, * p<0.1

[그림 3] 한국과 미국의 자살요인 다중집단 구조모형



***p<0.01, **p<0.05, *p<0.1 경로계수 표기: 한국경로계수(미국경로계수)

6) 본 연구의 구조모형은 그동안의 연구에서 제안된 스트레스가 건강생활실천요인(음주, 운동)을 매개하여 자살에 영향을 미칠 것이라는 스트레스 취약모형을 적용함

■ 자살요인 집단 간 구조모형 분석

○구조모형내 자살요인 변수 간의 인과관계에 있어 두 집단 사이에 유의미한 차이가 존재할 수 있어 모형 내 존재하는 모든 경로계수에 대해 각각 동일성 제약을 가한 모형을 기저모형과 비교

○스트레스에서 운동으로 가는 경로와 음주와 운동에서 자살로 가는 경로에서 집단간 유의미한 차이를 보이고 있음

– 한국이 미국보다 '스트레스→음주→자살'로 가는 경로와 '스트레스→운동→자살'로 가는 경로가 더 유의하게 강하게 나타남

〈표 3〉 기저모형과 경로계수에 동일성 제약을 가한 모형들 간의 집단 간 차이 비교

구분	χ^2	df	C.R[1)]	$\triangle\chi^2$
기저모형	2.061	2		
스트레스→음주	2.074	3	−.116	.013
스트레스→운동	30.979	3	5.529**	28.918
음주→자살	19.977	3	−4.328**	17.916
운동→자살	5.249	3	−1.803*	3.188
스트레스→자살	2.061	3	.000	0
모든경로제약	60.073	7		58.012

1) Critical ratios for differences

** p<0.05, * p<0.1

■ 자살요인 다중집단 구조모형의 매개효과 분석

○스트레스와 자살의 경로에 운동과 음주의 매개효과를 살펴보기 위해 효과분해를 실시한 결과 한국은 매개효과가 있으나 미국은 없는 것으로 나타났으며, 한국의 스트레스는 운동과 음주를 부분매개(partial mediation)하여 자살에 영향을 미치는 것으로 나타남

– 구글 검색트렌드에서 한국인이 스트레스를 경험할 경우 건강생활실천요인(음주, 운동)을 많이 찾게되고 이러한 건강생활실천요인이 자살 검색에 영향을 미치는 것으로 나타남

〈표 4〉 운동과 음주 매개변수의 집단 간 효과분해

경로	한국			미국		
	총효과	직접효과	간접효과[1)]	총효과	직접효과	간접효과[1)]
운동매개효과 스트레스→자살	.446	.277	.169*	.739	.706	.033
음주매개효과 스트레스→자살	.446	.380	.066**	.739	.737	.002

주: 1) Sobel Test: ** p<0.05 * p<0.1

4. 결론 및 빅 데이터 효율적 활용방안

■ 결론

○본 연구결과는 구글 검색트랜드의 빅 데이터 활용에 대한 사례에 불과하며 자살과 관련된 더 많은 빅 데이터를 활용할 경우, 자살을 사전에 예방할 수 있는 국가전략을 마련할 수 있을 것임

– 한국의 연도별 자살률은 구글의 자살 검색량과 비슷한 추세를 보이고 있음

– 연도별 자살 검색량의 결정요인에 대한 다층모형 분석결과 연도별 자살률은 자살 검색량에 유의하게 영향을 주는 것으로 나타남

– 다중집단 구조모형 분석결과 스트레스와 자살로 가는 경로에 건강생활실천요인이 미국은 없고 한국만 매개 효과가 있는 것으로 나타나, 한국인의 스트레스 해소를 위한 건강생활실천요인의 검색이 자살 검색에 영향을 주는 것으로 나타남

○본 연구결과에서 스트레스 검색량이 자살 검색량에 직접적인 영향을 주는 것으로 나타나 민간 검색포털이나 SNS에서 관련 키워드의 검색이나 Buzz 발생 시 스트레스와 자살충동을 감소시킬 수 있는 서비스의 제공이 요구됨

– 민간 검색포털이나 SNS에서 스트레스와 자살에 관련된 키워드의 검색이나 Buzz 발생 시, 이용자의 빅 데이터(연령층, 그동안의 검색패턴이나 Buzz 등)를 실시간으로 분석하여 징후가 예측되면 이용자에게 가장 적합한 스트레스 관리 프로그램을 팝업창이나 문자 메시지를 통하여 제공함으로써 자살충동을 예방하는 데 기여할 수 있을 것임

■ 보건복지 빅 데이터의 효율적 활용방안[7)]

○새로운 가치를 창출하고 예측가능한 보건복지 서비스를 제공할 수 있는 보건복지분야 빅 데이터의 효율적 활용을 위한 전략은 다음과 같음

– 보건복지 빅 데이터를 통합적으로 관리하기 위한 범 부처 차원의 조직(가칭: 보건복지 빅 데이터 관리 위원회)의 운영이 필요함

· 보건복지 빅 데이터는 보건복지부, 노동부, 여성가족부, 지식경제부, 통계청 등 많은 정부부처와 국민건강보험공단, 건강보험심사평가원, 식품의약품안전청, 국책연구기관 등 많은 공공기관에서 관리 · 운영되고 있어 각 기관에서 운영 중인 정보의 연계와 공유를 위해서는 범정부 차원의 조직이 필요함

– 보건복지 비 정형화된 빅 데이터를 관리하고 있는 민간기관과의 협조체제가 마련되어야 함

· 비 정형화된 보건복지 빅 데이터는 민간기관의 검색포털이나 SNS를 통해서 생산 · 저장되고 있어 민간기관과의 긴밀한 협조체계(가칭: 보건복지 빅 데이터 포럼)가 구축되어야 함

– 국가 차원의 Open API(Application Programming Interface)의 제공이 필요함

· 민간기관의 빅 데이터는 누구나 이용할 수 있기 때문에 전문가와 인프라만 있으면 분석이 가능하나 공공정보의 경우는 공개를 하지 않으면 활용이 불가함

7) '송태민. 보건복지 빅 데이터 효율적 활용방안. 보건복지포럼(2012. 11). 한국보건사회연구원' 의 내용을 재정리함

· 2012년 10월 기준으로 공유자원포털(www.data.go.kr)에서 공개되어 있는 API는 133종으로 이중 보건복지 분야의 API는 식품의약품안전청에서 제공하는 10종에 불과함

· 보건복지 빅 데이터의 공개는 관련기관과 빅 데이터 전문가의 참여로 정부와 국민이 필요로 하는 정보를 분류하고 공개대상 정보는 개인정보를 철저하게 보안하여 국가지식 플랫폼에 저장해야 함

– 보건복지 빅 데이터를 분석 처리할 수 있는 관련 기술의 개발이 필요함

· 스마트 시대에는 비 관계형, 비 정형 데이터의 저장과 분석, 클라우드 서비스의 확산, 시멘틱 검색서비스, 추론에 기반한 상황인식 서비스 등의 기술이 핵심임

· 관련부처와 협력하여 보건복지 분야 빅 데이터를 '수집→저장→분석→추론' 할 수 있는 기술개발은 물론 기술 표준화가 우선적으로 추진되어야 함

– 구조화되지 않은 대규모 데이터 속에서 숨겨진 정보를 찾아내는 데이터 사이언티스트(Data Scientist)의 인재 양성이 필요함

· 빅 데이터 시대에는 데이터를 관리하고 분석할 수 있는 인력이 매우 중요하며, 이미 글로벌 IT업체에서는 데이터 사이언티스트에 대한 인재 확보와 역량 강화를 추진함

– 보건복지 빅 데이터의 개인정보와 기밀정보에 대한 보안정책이 마련되어야 함

· 보건복지 빅 데이터는 개인에 대한 거의 모든 정보가 저장되어 있지만 아직 법 · 제도는 미비한 상황이며 논의조차 되지 못함

· 빅 데이터의 활용도 중요하지만 과다한 개인정보의 유출은 프라이버시 침해는 물론 사이버 인권침해나 범죄에 악용될 수 있음

· 빅 데이터로부터 개인을 보호하기 위해 가장 중요한 것은 특정 개인을 식별하지 못하도록 하는 익명화와 정보접근 및 정보처리에 대한 통제임

· 그러나, 정보접근 및 정보처리에 대한 통제를 강하게 하면 정보 활용이 활성화되지 않기 때문에 보건복지 빅 데이터의 '활용과 보호의 균형'에 대한 효과적인 정책이 마련되어져야 함

집필자 | 송태민(보건복지정보센터 연구위원) 문의 | 02-02-380-8201

발행인 | 최병호 **발행처** | 한국보건사회연구원
서울특별시 은평구 진흥로 235(122-705) | TEL 02)380-8000 | FAX 02)352-9129 | http://www.kihasa.re.kr
한국보건사회연구원 홈페이지의 발간자료에서 온라인으로도 이용하실 수 있습니다. http://www.kihasa.re.kr/html/jsp/publication/periodical/focus/list.jsp