Chapter 6
Integrating Non-clinical Data with EHRs

Yuan Lai, Edward Moseley, Francisco Salgueiro and David Stone

Take Home Messages

- Non-clinical factors make a significant contribution to an individual's health and providing this data to clinicians could inform context, counseling, and treatments.
- Data stewardship will be essential to protect confidential health information while still yielding the benefits of an integrated health system.

6.1 Introduction

The definition of "clinical" data is expanding, as a datum becomes clinical once it has a relation to a disease process. For example: the accessibility of one's home would classically be defined as non-clinical data, but in the context of a patient with a disability, this fact may become clinically relevant, and entered into the encounter note much like the patient's blood pressure and body temperature. However, even with this simple example, we can envision some of the problems with traditional non-clinical data being re-classified as clinical data, particularly due to its complexity.

6.2 Non-clinical Factors and Determinants of Health

Non-clinical factors are already significantly linked to health. Many public health policies focusing on transportation, recreation, food systems and community development are based on the relation between health and non-clinical determinants

The original version of this chapter was revised: A chapter author's name Edward Moseley was added. The erratum to this chapter is available at 10.1007/978-3-319-43742-2_30

MIT Critical Data, *Secondary Analysis of Electronic Health Records*,
DOI 10.1007/978-3-319-43742-2_6

such as behavioral, social and environmental factors [1]. Behavioral factors such as physical activity, diet, smoking and alcohol consumption are highly related to epidemic of obesity [2]. Some of this information, such as alcohol and tobacco use, is regularly documented by clinicians. Other information, such as dietary behaviors and physical activity, isn't typically captured, but may be tracked by new technology (such as wearable computers commonly referred to as "wearables") and integrated into electronic health records (EHRs). Such efforts may provide clinicians with additional context with which to counsel patients in an effort to increase their physical activity and reach a desired health outcome.

From a public health perspective, the same data obtained from these devices may be aggregated and used to guide decisions on public health policies. Continuing the prior example, proper amounts of physical activity will contribute to lower rates of mortality and chronic disease including coronary heart disease, hypertension, diabetes, breast cancer and depression across an entire population. Such data can be used to guide public health interventions in an evidence-based, cost-effective manner.

Both social and environmental factors are highly related to health. Social Determinants of Health (SDH) are non-clinical factors that affect the social and economic status of individuals and communities, including such items as their birthplace, living conditions, working conditions and demographic attributes [3]. Also included are social stressors such as crime, violence, and physical disorders, as well as others [4].

Environmental factors (i.e., air pollution, extreme weather, noise and poor indoor environmental quality) are highly related to an individual's health status. Densely built urban regions create air pollution, heat islands and high levels of noise, which have been implicated in causing or worsening a variety of health issues. For example, a study in New York City showed that asthma-related emergency admissions in youth from 5 to 17 years old were highly related to ambient ozone exposure. This annual NYC Community Health Survey also reveals that self-reported chronic health problems are related to extreme heat, suggesting that temperature can effect, or exacerbate, the symptoms of an individual's chronic illness. Social factors such as age and poverty levels also impact health. A study in New York City shows that fine particles ($PM_{2.5}$, a surrogate marker for pollution) attributable asthma hospital admissions are 4.5 times greater in high-poverty neighborhoods [5].

While outdoor environmental conditions merit public health attention, the average American spends only an hour of each day outdoors, and most individuals live, work and rest in an indoor environment, where other concerns reside. Poor indoor quality can cause building related illness and "sick building syndrome" (SBS)—where occupants experience acute health issues and discomfort, while no diagnosable illness can be readily identified [6]. Again in New York City, housing data was combined from multiple agencies in an effort to address indoor pollution concerns—using predictive analytics, the city was able to increase the rate of detection of buildings considered dangerous, as well as improve the timeliness in locating apartments with safety concerns or health hazards [7].

6.3 Increasing Data Availability

For many years scientists and researchers have had to deal with very limited available data to study behavioral, social and environmental factors that exist in cities, as well as the difficulty in evaluating their model with a large pool of urban data [8]. The big data revolution is bringing vast volumes of data and paradigmatic transformations to many industries within urban services and operations. This is particularly true in commerce, security and health care, as more data are systematically gathered, stored, and analyzed. The emergence of urban informatics also coincides with a transition from traditionally closed and fragmented data systems to more fully connected and open data networks that include mass communications, citizen involvement (e.g. social media), and informational flow [9].

In 2008, 3.3 billion of the world's inhabitants lived in cities, representing, for the first time in history the majority of the human population [10]. In 2014, 54 % of population lives in urban area and it is expected to increase to 66 % by 2050 [11]. With the growth of cities, there are rising concerns in public health circles regarding the impact of associated issues such as aging populations, high population densities, inadequate sanitation, environmental degradation, climate change factors, an increasing frequency of natural disasters, as well as current and looming resource shortages. A concomitantly large amount of information is required to plan and provide for the public health of these urban entities, as well as to prevent and react to adverse public events of all types (e.g. epidemiological, natural, criminal and politico-terroristic disasters).

The nature of the city as an agglomeration of inhabitants, physical objects and activities makes it a rich source of urban data. Today, billions of individuals are generating the digital data through their cellphones and use of the Internet including social networks. Hardware like global positioning systems (GPS) and other sensors are also becoming ubiquitous as they become more affordable, resulting in diverse types of data being collected in new and unique ways [12]. This is especially true in cities due to their massive populations, creating hotspots of data generation and hubs of information flow. Such extensive data availability may also provide the substrate for more statistically robust models across multiple disciplines.

An overview of the volume, variety, and format of open urban data is essential to further integration with electronic health records. As more cities begin building their informational infrastructure, the volume of city data increases rapidly. The majority of urban data are in tabular format with location-based information [8]. Data source and collection processes vary based on the nature of urban data. Passive sensors continuously collect environmental data such as temperature, air quality, solar radiation, and noise, and construct an urban sensing infrastructure along with ubiquitous computing [13]. There is also a large amount of city data generated by citizens such as service requests and complaints. Some pre-existing data, like those in the appropriate tabular format, are immediately ready for integration, while other data contained in more complex file types, like Portable

Document Format (PDF) or others, are more difficult to parse. This problem can be compounded if the data are encoded in uncommon character languages.

The fact that many non-clinical data, especially urban data, is geo-located enables clinicians to consider patient health within a broader view. Many environmental, social and behavioral factors link together spatially, and such spatial correlation is a key measurement in epidemiology, as it allows for the facilitation of data integration based on location. Connections and solutions become more visible by linking non-clinical data with EHR on a public health and city planning level. Recently, IBM announced that, by teaming supercomputer Watson's cognitive computing with data from CVS Health (a pharmacy chain with locations across the U.S.), we will have better predictions regarding the prevalence of chronic conditions such as heart disease and diabetes in different cities and locations [14].

6.4 Integration, Application and Calibration

In a summary of all cities in the United States that published open data sets as of 2013, it was found that greater than 75 % of datasets were prepared in tabular format [8]. Tabular data is most amenable for automated integration, as it is already in the final format prior to being integrated into most relational databases (as long as the dataset contains a meaningful attribute, or variable, with which to relate to other data entries). Furthermore, data integration occurs most easily when the dataset is "tidy", or follows the rule of "one observation per row and one variable per column." Any data manipulation process resulting in a dataset that is aggregated or summarized could remove a great deal of utility from that data [15].

For instance, a table that is familiar within one working environment may not be easily decipherable to another individual and may be nearly impossible for a machine to parse without proper context given for what is within the table. An example could be a table of blood pressure over time and in different locations for a number of patients, which may look like (Table 6.1).

Here we see two patients, Patient 1 and Patient 2, presenting to two locations, Random and Randomly, RA, on two different dates. While this table may be easily read by someone familiar with the format, such that an individual would understand that Patient 1 on the 1st of January, 2015, presented to a healthcare setting in Random, RA with a systolic blood pressure of 130 mmHg and a diastolic pressure of 75 mmHg, it may be rather difficult to manipulate these data to a tidy format without understanding the context of the table.

Table 6.1 Example of a table requiring proper context to read

Patient blood pressure chart	Random, RA		Randomly, RA	
	1-Jan-15	7-Jan-15	1-Jan-15	7-Jan-15
Patient 1	130/75	139/83	141/77	146/82
Patient 2	158/95	151/91	150/81	141/84

If this table were to be manipulated in a manner that would make it easily analyzed by a machine (as well as other individuals without requiring an explanation of the context), it would follow the rule of one column per variable and one row per observation, as below (Table 6.2).

There are further limitations imparted due to data resolutions, which refers to the detail level of data in space, time or theme, especially the spatial dimension of the data [16]. Examples include: MM/DD/YY time formats compared to YYYY; or zip codes compared to geographic coordinates. Even with these limitations, one may still be able to draw relevant information from these spatial and temporal data.

One method to provide spatial orientation to a clinical encounter has recently been adopted by the administrators of the Medical Information Mart for Intensive Care (MIMIC) database, which currently contains data from over 37,000 intensive care unit admissions [17]. Researchers utilize the United States Zip Code system to approximate the patients' area of residence. This method reports the first three digits of the patient's zip code, while omitting the last two digits [18]. The first three digits of a zip code contain two pieces of information: the first integer in the code refers to a number of states, the following two integers refer to a U.S. Postal Service Sectional Center Facility, through which the mail for that state's counties is processed [19]. The first three digits of the zip code are sufficient to find all other zip codes serviced by the Sectional Center Facility, and population level data of many types are available by zip code as per the U.S. Government's census [20].

Table 6.2 A tidy dataset that contains a readily machine-readable format of the data in Table 6.1

Patient ID	Place	Date (MM/DD/YYYY)	Pressure (mmHg)	Cycle
1	Random, RA	1/1/2015	130	Systole
1	Random, RA	1/1/2015	75	Diastole
1	Random, RA	1/7/2015	139	Systole
1	Random, RA	1/7/2015	83	Diastole
1	Randomly, RA	1/1/2015	141	Systole
1	Randomly, RA	1/1/2015	77	Diastole
1	Randomly, RA	1/7/2015	146	Systole
1	Randomly, RA	1/7/2015	82	Diastole
2	Random, RA	1/1/2015	158	Systole
2	Random, RA	1/1/2015	95	Diastole
2	Random, RA	1/7/2015	151	Systole
2	Random, RA	1/7/2015	91	Diastole
2	Randomly, RA	1/1/2015	150	Systole
2	Randomly, RA	1/1/2015	81	Diastole
2	Randomly, RA	1/7/2015	141	Systole
2	Randomly, RA	1/7/2015	84	Diastole

Connections and solutions become more visible by linking non-clinical data with EHRs on a public health and city planning level. Although many previous studies show the correlation between air pollution and asthma, it is only recently individuals became able to trace PM2.5, SO_2 and Nickel (Ni) in the air back to the generators in buildings with aged boilers and heating systems, which is due in large part to increasing data collection and integration across multiple agencies and disciplines [21]. As studies reveal additional links between our environment and pathological processes, our ability to address potential health threats will be limited by our ability to measure these environmental factors in sufficient resolution to be able to apply it to patient level, creating truly personalized medicine.

For instance, two variables, commonly captured in many observations, are geo-spatiality and temporality. Since all actions share these conditions, integration is possible among a variety of data otherwise loosely utilized in the clinical encounter. When engaged in an encounter, a clinician can determine, from data collected during the examination and history taking, the precise location of the patient over a particular period of time within some spatial resolution. As a case example, a patient may present with an inflammatory process of the respiratory tract. The individual may live in random, RA, and work as an administrator in Randomly, RA; one can plot these variables over time, and separate them to represent both the individuals' work and home environment—as well as other travel (Fig. 6.1).

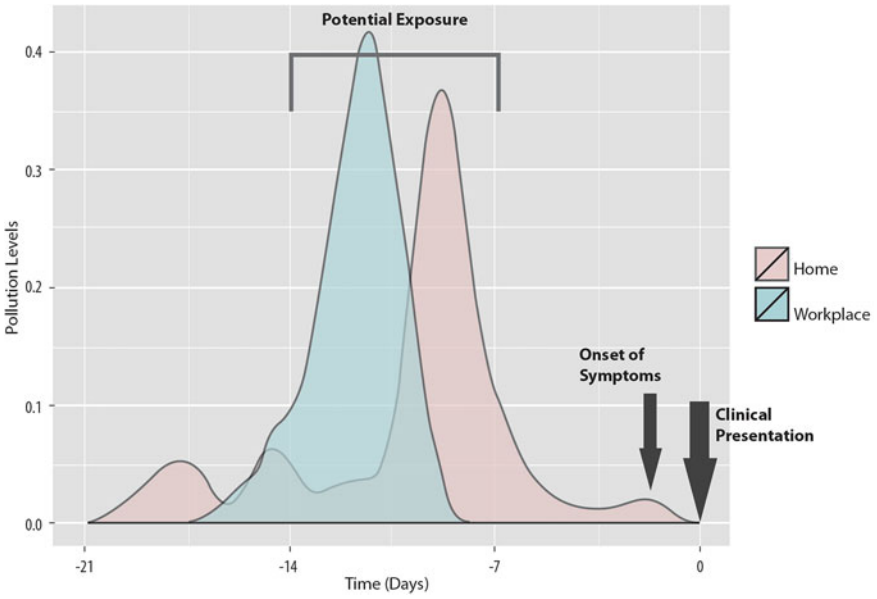

Fig. 6.1 Example of pollution levels over time for a patient's "work" and "home" environment with approximate labels that may provide clinically relevant decision support

This same method may be applied to other variables that could be determined to have statistical correlates of significance during the timeframe prior to the onset of symptoms and then the clinical encounter.

With the increasing availability of information technology, there is less need for centralized information networks, and the opportunity is open for the individual to participate in data collection, creating virtual sensor networks of environmental and disease measurement. Mobile and social web have created powerful opportunities for urban informatics and disaster planning particularly in public health surveillance and crisis response [13]. There are geo-located mobile crowdsourcing applications such as Health Map's Outbreaks Near Me [22] and Sickweather [23] collecting data on a real-time social network.

In the 2014 Ebola Virus Disease outbreak, self-reporting and close contact reporting was essential to create accurate disease outbreak maps [24]. The emergence of wearables is pushing both EHR manufacturers to develop frameworks that integrate data from wearable devices, and third party companies to provide cloud storage and integration of data from different wearables for greater analytic power.

Attention and investment in digital health and digital cities continues to grow rapidly. In digital health care, investors' funding has soared from $1.1 Billion in 2011 to $5.0 Billion in 2014, and big data analytics ranks as the #1 most active subsector of digital healthcare startups in both amount of investment and number of deals [25]. Integration will be a long process requiring digital capabilities, new policies, collaboration between the public and private sectors, and innovations from both industry leaders and research institutions [26]. Yet we believe with more interdisciplinary collaborations in data mining and analytics, we will gain new knowledge on the health-associated non-clinical factors and indicators of disease outcomes [27]. Furthermore, such integration creates a feedback loop, pushing cities to collect better and larger amounts of data. Integrating non-clinical information into health records remains challenging. Ideally the information obtained from the patient would flow into the larger urban pool and vice versa. Challenges remain on protecting confidentiality at a single patient level and determining applicability of macroscopic data to the single patient.

6.5 A Well-Connected Empowerment

Disease processes can result and be modified by interactions of the patient and his or her environment. Understanding this environment is of importance to clinicians, hospitals, public health policy makers and patients themselves. With this information we can preempt patients at risk for disease (primary prevention), act earlier in minimizing morbidity from disease (secondary prevention) and optimize therapeutics.

A good example of the use of non-clinical data for disease prevention is the use of geographical based information systems (GIS) for preemptive screening of

populations at risk for sexually transmitted diseases (STDs). Geographical information systems are used for STD surveillance in about 50 % of state STD surveillance programs is the U.S. [28]. In Baltimore (Maryland, U.S.) a GIS based study identified core groups of repeat gonococcal (an STD) infection that showed geographical clustering [29]. The authors hinted at the possibility of increased yield when directing prevention to geographically restricted populations.

A logical next step is the interaction between public health authority systems and electronic medical records. As de-identified geographical health information becomes publically available, an electronic medical record would be able to download this information from the cloud, apply it to the patient's zip code, sex, age and sexual preference (if documented) and warn/cue the clinician that would decide if an intervention is required based on a calculated risk to acquire a STD.

6.6 Conclusion

Good data stewardship will be essential for protecting confidential health information from unintended and illegal disclosure. For patients, the idea of increasing empowerment in their health is essential [8]. Increasing sensor application and data visualization make our own behavior and surroundings more visible and tangible, and alert us about potential environmental risks. More importantly, it will help us to better understand and gain power over our own lives.

The dichotomy of addressing population health versus individual health must be addressed. Researchers should ask: what information is relevant to the target which I'm addressing, and what data do we feed from this patient's record into the public health realm? The corollary to that question is: how can we balance the individual's right to privacy with the benefit of non-clinical data applicable to the individual and to the large populations? Finally: how can we create systems that select relevant data from a single patient and present it to the clinician in a population-health context? In this chapter, we have attempted to provide an overview of the potential use of traditionally non-clinical data in electronic health records, in addition to mapping some of the pitfalls and strategies to using such data, as well as highlighting practical examples of the use of these data in a clinical environment.

References

1. Barton H, Grant M (2013) Urban planning for health cities, a review of the progress of the european healthy cities program. J Urban Health Bull NY Acad Med 90:129–141
2. Badland HM, Schofield GM, Witten K, Schluter PJ, Mavoa S, Kearns RA, Hinckson EA, Oliver M, Kaiwai H, Jensen VG, Ergler C, McGrath L, McPhee J (2009) Understanding the relationship between activity and neighborhoods (URBAN) study: research design and methodology. BMC Pub Health 9:244
3. Osypuk TL, Joshi P, Geronimo K, Acevedo-Garcia D (2014) Do social and economic policies influence health? Rev Curr Epidemiol Rep 1:149–164
4. Shmool JLC, Kubzansky LD, Newman OD, Spengler J, Shepard P, Clougherty JE (2014) Social stressors and air pollution across New York City communities: a spatial approach for assessing correlations among multiple exposures. Environ Health 13:91
5. Kheirbek I, Wheeler K, Walters S, Kass D, Matte T (2013) PM2.5 and ozone health impacts and disparities in New York City: sensitivity to spatial and temporal resolution. Air Qual Atmos Health 6:473–486
6. Indoor Air Facts No. 4 sick building syndrome. United States Environmental Protection Agency, Research and Development (MD-56) (1991)
7. Goldstein B, Dyson L (2013) Beyond transparency: open data and the future of civic innovation. Code for America Press, San Francisco
8. Barbosa L, Pham K, Silva C, Vieira MR, Freire J (2014) Structured open urban data: understanding the landscape. Big Data 2:144–154
9. Shane DG (2011) Urban design since 1945—a global perspective. Wiley, New York, p 284
10. National Intelligence Council (2012) Global trends 2030: alternative worlds. National Intelligence Council
11. World Urbanization Prospects, United Nations (2014)
12. Goldsmith S, Crawford S (2014) The responsive city: engaging communities through data-smart governance. Wiley, New York
13. Boulos M, Resch B, Crowley D, Breslin J, Sohn G, Burtner R, Pike W, Jezierski E, Chuang K (2011) Crowdsourcing, citizen sensing and sensor web technologies for public and environmental health surveillance and crisis management: trends, OGC standards and application examples. Int J Health Geographic 10:67
14. McMullan T. Dr Watson: IBM plans to use Big Data to manage diabetes and obesity. URL: http://www.alphr.com/life-culture/1001303/dr-watson-ibm-plans-to-use-big-data-to-manage-diabetes-and-obesity
15. Wickham H (2014) Tidy data. J Stat Softw 10:59
16. Haining R (2004) Spatial data analysis—theory and practice. Cambridge University Press, Cambridge, p 67
17. MIMIC II Databases. Available from: http://physionet.org/mimic2. Accessed 02 Aug 2015
18. Massachusetts Institute of Technology, Laboratory of Computational Physiology. mimic2 v3.0 D_PATIENTS table. URL: https://github.com/mimic2/v3.0/blob/ad9c045a5a778c6eb283bdad310594484cca873c/_posts/2015-04-22-dpatients.md. Accessed 02 Aug 2015 (Archived by WebCite® at http://www.webcitation.org/6aUNzhW6g)
19. http://pe.usps.com/businessmail101/glossary.htm
20. http://factfinder.census.gov/
21. Jeffery N, McKelvey W, Matte T (2015) using tracking infrastructure to support public health programs, policies, and emergency response in New York City. Pub Health Manag Pract 21(2 Supp):S102–S106
22. http://www.healthmap.org/outbreaksnearme/
23. http://www.sickweather.com

24. Kouadio KI, Clement P, Bolongei J, Tamba A, Gasasira AN, Warsame A, Okeibunor JC, Ota MO, Tamba B, Gumede N, Shaba K, Poy A, Salla M, Mihigo R, Nshimirimana D (2015) Epidemiological and surveillance response to Ebola virus disease outbreak in Lofa County, Liberia (Mar–Sept 2014); lessons learned, edn 1. PLOS Currents Outbreaks. 6 May 2015. doi: 10.1371/currents.outbreaks.9681514e450dc8d19d47e1724d2553a5
25. The re-imagination of healthcare. StartUp Health Insights. www.startuphealth.com/insights
26. Ericsson Networked Society City Index (2014)
27. Corti B, Badland H, Mavoa S, Turrell G, Bull F, Boruff B, Pettit C, Bauman A, Hooper P, Villanueva K, Burt T, Feng X, Learnihan V, Davey R, Grenfell R, Thackway S (2014) Reconnecting urban planning with health: a protocol for the development and validation of national livability indicators associated with non-communicable disease risk behaviors and health outcomes. Pub Health Res Pract 25(1):e2511405
28. Bissette JM, Stover JA, Newman LM, Delcher PC, Bernstein KT, Matthews L (2009) Assessment of geographic information systems and data confidentiality guidelines in STD programs. Pub Health Rep 124(Suppl 2):58–64
29. Bernstein TK, Curriero FC, Jennings JM et al (2004) Defining core gonorrhea transmission utilizing spatial data. Am J Epidemiol 160:51–58