**FRONT MATTER**

**Title**
Surveillance of non-communicable diseases: opportunities in the era of big data

**Short title**
Surveillance of NCDs and big data

**Authors**
Pengfei Li,1,2† Lin Ma,3† Jue Liu,4 Luxia Zhang1,2*

**Affiliations**
1. Advanced Institute of Information Technology, Peking University, Hangzhou 311215, China
2. National Institute of Health Data Science, Peking University, Beijing 100191, China
3. Peking University Health Science Center, Beijing, 100191, China
4. Department of Epidemiology and Biostatistics, School of Public Health, Peking University, Beijing 100191, China.
†These authors contributed equally to this work
*Corresponding author. Email: zhanglx@bjmu.edu.cn

**MAIN TEXT**

Non-communicable diseases (NCDs) are the leading cause of mortality, accounting for 70% of deaths worldwide, and have become one of the major challenges for human sustainable development in the 21st century.[1] The pandemic of coronavirus disease 2019 (COVID-19) interrupted NCDs services in 75% of the countries—public health campaigns and NCD surveillance efforts gave way to pandemic control,[2] which further imposes challenges to fight NCDs. Among the key components of strategy for the prevention and control of NCDs, surveillance is a crucial one to track and monitor the major risk factors and intervention effects, which is defined as “the ongoing systematic collection, analysis, and interpretation of health data essential to the planning, implementation, and evaluation of public health practice, closely integrated with timely dissemination of these data to those who need to know”[3, 4]. The capacity to undertake surveillance varied substantially and was inadequate in many low-income and middle-income countries. It is hoped that integration of multiple existing data sources into a comprehensive NCDs surveillance framework could enhance the NCDs surveillance efforts.[5]

1. **The current NCDs surveillance system**

To guide the implementation of medical and public health measures to control the diseases, Alwan et al proposed the national surveillance framework for NCDs that encompassed key risk factors, outcomes, health-system response interventions, and health-system capacity.[6] Considering the need of monitoring incidence and the emphasis on representativeness, the periodic population-based survey is one of the most important components of the surveillance system.

The National Health and Nutrition Examination Survey (NHANES) in the United States initiated in the 1960s is one of the early endeavors to monitor the health and nutritional status of nation-wide population. Starting in 1999, NHANES became a continuous program focusing on a variety of health and nutrition measurements each year.[7] Similar
population-based surveys have been launched in low-income and middle-income countries (LMICs) with a high burden of NCDs. For example, in China, the nutrition and chronic diseases surveys have been combined since 2002 and the resulting China National Nutrition and Health Surveys provide information of “how the social and economic transformation of Chinese society is affecting the health and nutritional status of its population”.

Population-based surveys could be extremely manpower, time, and material resources-consuming, limiting NCDs surveillance capacity, especially in LMLCs. For example, among 22 LMLCs with a high burden of NCDs, only 50% of them have reported population-based information on risk factors for NCDs. Even for countries with risk factors information, concerns exist regarding the accuracy, quality, and standardization of data. Furthermore, the composition of global burden of diseases has displayed a dynamic trend and the current surveillance system might not have the resilience to include emerging NCDs in a timely manner. For instance, chronic kidney disease (CKD) has only been recognized as an important public health issue worldwide in the past 20 years while the surveillance programs (especially national programs initiated by the government) for CKD are extremely limited. In addition, during major events such as the COVID-19 outbreak, risk behaviors affecting health could change rapidly. As a result, the current periodic survey-based system might not be adequate to identify those acute changes. The resource consumption, incomplete reporting, and time lags of existing surveillance systems might damage the ability of governments and health organizations to capture NCD trends and manage unmet needs.

2. Big data and NCDs surveillance systems

The digitalization in medicine and the advent of big data analytics have introduced novel opportunities for NCDs surveillance. The widespread use of electronic health records (EHR), the accumulating administrative and claim data at national level, and other types of datasets have enabled us to pursue solutions to population health issues previously thought impossible. Instead of extrapolating from the data obtained from samples at high costs to make inferences about a population of national level, it's possible to use EHR data across health organizations at the national population level to provide a real-world picture. The practice-driven and longitudinal nature of real-world data enable the inclusion of multiple risk factors, outcomes, and intervention effect evaluation, which are all important elements of NCDs surveillance. Furthermore, integrating and analyzing enriched data resources have introduced the feasibility of quick identification and response to changing situations associated with NCD surveillance. As an illustration of the value of multiple data resources, the surveillance of CKD as an emergent NCD is used as an example in the followings.

After the concept of CKD was defined in 2002 and existing data revealed its public health importance, the Centers for Disease Control and Prevention in the United States launched the CKD initiative to establish the first comprehensive CKD Surveillance System in 2006. The approach is to leverage existing data from a wide range of data sources, including data from both healthcare systems and data from existing surveillance programs. Through the integration of various data sources available, the CKD Surveillance System can provide detailed information in support of CKD control in many important aspects.
The similar approach could also be utilized in developing countries with limited resources and capacity. In China, the prevalence of CKD is reported to be 10.8% and imposes a substantial burden on the healthcare system, while CKD was not included in the existing government-initiated surveillance system. Under the national strategy to promote the application of big data, an initiative entitled China Kidney Disease Network (CK-NET) was launched in 2014. CK-NET integrates and analyzes different data sources—administrative data, regional HER, research data, and real-world data, being recognized as the most comprehensive CKD surveillance system and “an important benchmark for kidney disease surveillance in China”.

### 3. The future of NCDs surveillance

Besides big data and data analytics, many cutting-edge technologies could be used to collect data complementing existing NCDs surveillance data. As home-based wireless devices, apps, and wearables technology mature in recent years, the real-time health data generated by patients can be collected, which could capture health behaviors including patient engagement. During the pandemic of COVID-19, the cutting-edge technologies are pivotal in infection surveillance. Besides wearable health monitoring sensors, virtual care technologies and Internet-of-Things could be integrated to develop smart disease surveillance systems to prevent, diagnose, and treat COVID-19. Such kinds of systems could also be instrumental for NCDs surveillance.

Furthermore, it is increasingly recognized that health is ineluctably linked to social, environmental, and economic factors, hence including behavioral, environmental, network, and community data into the surveillance of NCDs could lead to opportunities for interventions aimed at improving population health. For example, accumulating evidence suggests that worsening economic outcomes may be a primary contributor to negative health trends among working-age US adults with low income and less education. In addition, virtual digital trails such as mining social media data provide the opportunity to evaluate self-reported NCD-related attitudes and behaviors. Research has shown conveying anger and fatigue on Twitter was associated with an increased risk of heart disease. With the increased complexity of collected data in the surveillance system, tools from other sectors could also be utilized to develop innovative approaches for NCDs surveillance. For example, using spatial analyses and geographical information systems provided a geographic perspective to explore health disparities and access to care in patients with head and neck cancer.

In summary, the outburst of data in both the medical area and beyond, the development of data analytics and other cutting-edge technologies, and the emergence of data-driven paradigms are transforming NCDs surveillance. Although extremely promising, there are still enormous legal, ethical, political, and technical challenges to overcome. In addition, transdisciplinary efforts are unprecedentedly important in the era of big data.

### Funding

This work was supported by grants from the National Natural Science Foundation of China [grant number 72125009].

### Authors’ Contributions
PL and LZ conceived and designed the study. PL and LM drafted the manuscript. JL provided epidemic domain expertise. PL and LM contributed equally to this work. All authors critically revised the manuscript for important intellectual content. All authors approved the final version for submission.

**Conflicts of Interest**

All authors have completed the ICMJE uniform disclosure form. The authors declare that there is no conflict of interest regarding the publication of this article.

**References**


