Application of Big Data analysis in gastrointestinal research

Ka-Shing Cheung, Wai K Leung, Wai-Kay Seto

ORCID number: Ka-Shing Cheung (0000-0002-4838-378X); Wai K Leung (0000-0002-5993-1059); Wai-Kay Seto (0000-0002-9012-313X).

Author contributions: All authors contributed equally to this paper with literature review and analysis, drafting and critical revision and editing, and approval of the final version of this article.

Conflict-of-interest statement: WKL has received an honorarium for attending advisory board meetings of Boehringer Ingelheim and Takeda. WKS received honorarium for attending advisory board meetings of AbbVie, Celltrion and Gilead; speaker fees from AbbVie, AstraZeneca, Eisai, Gilead and Ipsen; and research funding from Gilead.

Open-Access: This article is an open-access article that was selected by an in-house editor and fully peer-reviewed by external reviewers. It is distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/

Manuscript source: Invited manuscript

Received: March 12, 2019
Peer-review started: March 13, 2019
First decision: April 11, 2019
Revised: April 14, 2019
Accepted: April 29, 2019

Abstract

Big Data, which are characterized by certain unique traits like volume, velocity and value, have revolutionized the research of multiple fields including medicine. Big Data in health care are defined as large datasets that are collected routinely or automatically, and stored electronically. With the rapidly expanding volume of health data collection, it is envisioned that the Big Data approach can improve not only individual health, but also the performance of health care systems. The application of Big Data analysis in the field of gastroenterology and hepatology research has also opened new research approaches. While it retains most of the advantages and avoids some of the disadvantages of traditional observational studies (case-control and prospective cohort studies), it allows for phenomapping of disease heterogeneity, enhancement of drug safety, as well as development of precision medicine, prediction models and personalized treatment. Unlike randomized controlled trials, it reflects the real-world situation and studies patients who are often under-represented in randomized controlled trials. However, residual and/or unmeasured confounding remains a major concern, which requires meticulous study design and various statistical adjustment methods. Other potential drawbacks include data validity, missing data, incomplete data capture due to the unavailability of diagnosis codes for certain clinical situations, and individual privacy. With continuous technological advances, some of the current limitations with Big Data may be further minimized. This review will illustrate the use of Big Data research on gastrointestinal and liver diseases using recently published examples.

Key words: Healthcare dataset; Epidemiology; Gastric cancer; Inflammatory bowel disease; Colorectal cancer; Hepatocellular carcinoma; Gastrointestinal bleeding

©The Author(s) 2019. Published by Baishideng Publishing Group Inc. All rights reserved.
Core tip: Digital collection and storage of data has led to the generation of Big Data. Big Data analysis in the field of gastroenterology and hepatology allows for phenomapping due to disease heterogeneity (e.g., inflammatory bowel disease, gastrointestinal and liver cancers) and hence the development of precision medicine, enhances in drug safety and faster drug discovery. It has also revolutionized clinical study approaches. Although there are still limitations to Big Data approaches, some of them may be further minimized with continuous technological advances.

Citation: Cheung KS, Leung WK, Seto WK. Application of Big Data analysis in gastrointestinal research. World J Gastroenterol 2019; 25(24): 2990-3008
URL: https://www.wjgnet.com/1007-9327/full/v25/i24/2990.htm
DOI: https://dx.doi.org/10.3748/wjg.v25.i24.2990

INTRODUCTION

The etymology of “Big Data” can be dated back to the 1990s, and this term has become popular after John Mashey, the then chief scientist at Silicon Graphics[1]. Datasets are exponentially expanding every day, fed with a wide array of sources like mobile communications, websites, social media/crowdsourcing, sensors, cameras/lasers, transaction process-generated data (e.g., sales queries, purchases), administrative, scientific experiments, science computing, and industrial manufacturing. The application of Big Data analysis has proven successful in many fields. Technology giants (e.g., Amazon, Apple, Google) have boosted sales and increased revenue by means of Big Data approaches[3]. It has also been adopted as part of the electoral strategies in political campaigns[4].

There is currently no consensus on the definition of Big Data, but the characteristics pertinent to the process of collection, storage, processing and analysis of these data helps to forge Big Data as a more tangible term. It was first described by Doug Laney in 2001 that Big Data possessed 3Vs: Volume (storage space necessary for data recording and storage), Velocity (speed of data generation and transformation) and Variety (various data sources)[5]. Since then, many other traits to define Big Data have been proposed, including veracity, value, exhaustivity (n=all), fine-grained resolution, indexicality, relationality, extensionality, scalability, and variability[6].

BIG DATA RESEARCH IN GASTROENTEROLOGY AND HEPATOLOGY

The digitalization of nearly every aspect of daily life has made no exception in the field of healthcare, with the importance of Big Data application being increasingly recognised and advocated in recent years. While there are various definitions of Big Data outside of the medical field, the specific definition with respect to health has only been proposed in recent years. According to the report produced under the third Health Programme (2014-2020) from the Consumer, Health, Agriculture and Food Executive Agency mandated by the European Commission[6], Big Data in Health are defined as large datasets that are collected routinely or automatically, and stored electronically. It merges existing databases and is reusable (i.e., multipurpose data that are not intended for a specific study), with the aim of improving health and health system performance. A further supplement is the scale and complexity of the data that mandates dedicated analytical and statistical approaches[7]. Such large volume and scale of Big Data arise not only from the number of subjects included, but also the diversity of variables from different domains (clinical, lifestyle, socioeconomic, environmental, biological and omics) at several time points. The estimated healthcare volume of 153 exabytes (10^18) in 2014 is projected to hit 2,300 exabytes by 2020[8,9].

Big Data in Health relies on a wealth of sources: Administrative databases, insurance claims, electronic health records, cohort study data, clinical trial data, pharmaceutical data, medical images, biometric data, biomarker data, omics data (e.g., genomics, proteomics, metabolomics, microbiomics), social media (e.g., Facebook, Twitter), income statistics, environmental databases, mobile applications, e-Health tools, and telemedicine (diagnosis and management at a distance, particularly by means of the internet, mobile phone applications and wearable devices)[9]. The
importance of “data fusion” therefore relies on the systematic linking of datasets from different sources to add values and new insights, enabling the analysis of health data from different perspectives (individual, group, social, economic and environmental factors) across different regions or nations.

Disease entities in the field of gastroenterology and hepatology are often heterogeneous [e.g., malignancy, inflammatory bowel disease (IBD)] with a wide range of clinical phenotypes (e.g., age of onset, severity, natural course of disease, association with other diseases, treatment response). Big Data analysis allows for the subclassification of a disease entity into distinct subgroups (i.e., phenomapping), which enhances understanding of disease pathogenesis, as well as the development of more precise predictive models of disease outcomes. The use of only clinical and laboratory data (as in traditional clinical research) in predicting disease course, outcome and treatment response may not achieve a high accuracy\(^{[1]}\). Similarly, although genome-wide association studies (commonly known as GWAS) and identification of single nucleotide variants have linked particular disease phenotypes to genetic defects, most genetic variants have a small impact on disease risk, behaviour and treatment response\(^{[2]}\). This inaccurate differentiation has led to the unnecessary use of therapeutics (which are sometimes costly with undesirable side effects) in many patients (e.g., biologics in IBD patients). It therefore appears that only by considering the complex interactions between genetic, lifestyle, environmental factors, and previously unconsidered factors (e.g., omics) in Big Data approaches can a reliable predictive diagnostic model be developed, which ultimately guides a targeted approach for selecting treatment regimens for individual patients (i.e., precision or personalized medicine)\(^{[3,12]}\).

Apart from phenomapping and precision medicine, other important implications of Big Data approaches are drug discovery and safety. Drug research and development (R and D) is an expensive and lengthy process, with each drug approval costing $3.2-32.3 billion US dollars\(^{[4]}\). Many of the trial drugs have proven futile or harmful in early or even late stages of the development (e.g., secukinumab in Crohn’s disease\(^{[5]}\)). Even for drugs proven to be beneficial, they may only work in certain subgroups of patients. The heterogeneity of therapeutic outcomes is again likely multifactorial. Precision medicine from Big Data approaches will help pharmaceutical companies predict drug action and prioritize drug targets on a specific group of patients\(^{[6]}\). This ensures a cost-effective approach in developing new therapeutics with a lower chance of futility.

Recently, “drug repositioning” or “drug repurposing” has been advocated, in which currently approved drugs are explored for other indications of gastrointestinal and hepatic diseases. However, to make sense of the large-scale genomic and phenotypic data, advanced data processing and analysis is an indispensable element, hence giving rise to the term “computational drug repositioning or repurposing”\(^{[7]}\). This involves a process of various computational repositioning strategies utilizing different available data sources, computational repositioning approaches (e.g., machine learning, network analysis, text mining and semantic inference), followed by validation via both computational (electronic health records) and experimental methods (in vitro and in vivo models). Applicable disease areas include oncology [e.g., hepatocellular carcinoma (HCC)]\(^{[8,9]}\), infectious diseases, and personalized medicine, just to name a few. New indications of existing medications constituted 20% of 84 drugs products introduced to the market in 2013\(^{[10]}\). Drug repositioning is expected to play an increasingly important role in drug discovery for gastrointestinal and liver diseases.

With regards to drug safety, monitoring currently relies on data from randomized controlled trials (RCTs) or post-marketing studies. However, RCTs may be underpowered to detect rare but important side effects, and fail to capture adverse effects that only manifest beyond the designed follow-up time (e.g., malignancy). Post-marketing studies based on registries are resource-intensive in terms of cost and time, and the safety profile of a drug can only be depicted several years after marketing. The application of text mining, the computational process of extracting meaningful information from unstructured text, has proven useful to improve pharmacovigilance (e.g., arthralgia in vedolizumab users in IBD\(^{[11]}\)). The sources are not limited to medical literature and clinical notes, but also product labelling, social media and web search logs\(^{[12,21]}\).

**ADVANTAGES AND SHORTCOMINGS OF BIG DATA APPROACHES**

In healthcare research, RCT is regarded as the gold standard to investigate the
causality between exposure and the outcome of interest. Randomization balances prognostic factors across intervention and control groups. It eliminates both measured and unmeasured confounding, making the establishment of causality possible. However, it is resource-intensive to conduct RCTs in terms of money, manpower and time. It is difficult to study rare events (e.g., cancer, death) or long-term effects. Due to the stringent inclusion and exclusion criteria, as well as differential levels of care and follow-up in a clinical trial setting, results from RCTs may not reflect real-life situations, and may not be generalizable to other populations. Finally, effects of harmful exposure cannot be studied due to ethical concerns.

To circumvent these shortcomings of RCTs, observational studies are alternatives. Case-control studies are cheaper and quicker to conduct, and can study multiple risk factors of rare diseases, as well as potentially harmful exposure that is otherwise impossible in RCTs. On the other hand, prospective cohort studies can investigate multiple exposures and outcomes, effects of rare exposure, as well as potentially harmful exposure. Nonetheless, it is difficult to study rare exposures in case-control studies, as well as rare diseases or long-term effects in prospective cohort studies. It is also impossible and unethical to prospectively follow the natural history of chronic diseases and its complications without appropriate interventions\(^\text{28}\). In addition, for both study designs, multiple biases (e.g., reverse causality, selection bias, interviewer bias, recall bias) can exist, and confounding, whether measured or unmeasured, is always possible.

The application of Big Data analysis in healthcare research has revolutionized clinical study approaches. Clinical studies making use of these datasets usually belong to either retrospective cohort studies (non-concurrent/historical cohort studies) or nested case-control studies. As the clinical data are readily available without delays, and easily retrieved from the electronic storage system, a multitude of risk factors can be included to analyse the outcome. It also enables the study of rare exposures, rare events and long-term effects within a relatively short period of time. Resources are much less than that required for prospective cohort study design, except for dedicated manpower with the aid of high-performance computers and software, e.g., R, Software for Statistics and Data Science, Statistics Analysis System, Python. In essence, it retains most of the advantages while avoiding some of the disadvantages of case-control and prospective cohort studies. Unlike RCTs, it reflects the real-world efficacy, and studies patients who are often under-represented in or completely excluded from RCTs (e.g., the elderly, pregnant women). Furthermore, the huge sample size of Big Data permits subgroup analysis to investigate interactions between different variables with the outcome of interest without sacrificing statistical power. It enables the investigation of varying factors (i.e., division of the follow-up duration into different segments) on the association between exposure and outcome, given a sufficiently long observation period (in terms of years or decades) and sample size. It also allows for multiple sensitivity analyses by including certain sub-cohorts, modifying definitions of exposure (e.g., duration of drug use), or different statistical methods to prove the robustness of study results. A reliable capture of small variations in incidence or flares of a disease according to temporal variations also heavily depend on the sample size. In the most ideal situation of \(n = \text{all}\), selection bias will no longer be a concern.

However, it should be acknowledged that without randomization, residual and/or unmeasured confounding remains a concern in Big Data research. As such, one may argue that causality cannot be established. The inclusion of RCT datasets with the extensive collection of data and outcomes for trial participants or linkage with other data sources may partly address this issue\(^\text{29}\). The possibility of causality can also be strengthened via the fulfilment of the Bradford Hill criteria\(^\text{30}\). Second, data validity concerning the accuracy of diagnosis codes (e.g., International Classification of Diseases) in electronic databases has been challenged\(^\text{31}\). In addition, milder disease tends to be omitted in the presence of more serious disease, and hence the absence of a diagnosis code may not signify the absence of that particular disease\(^\text{32}\). For instance, depression, which is often not coded among the elderly with other serious medical diseases, may be paradoxically associated with reduced mortality. To a certain extent, data validity can be verified through validating the diagnosis codes by cross referencing the actual diagnosis of a subset of patients in the medical records.

Third, missing data can potentially bias the result via a differential misclassification bias. There are different remedies, although the use of multiple imputation is preferred, which involves constructing a certain number of complete datasets (e.g., \(n = 50\)) by imputing the missing variables based on the logistic regression model\(^\text{33}\). Nonetheless, missing data with differential misclassifications are not a major problem in Big Data health research, as diagnosis codes are recorded by healthcare professionals, with other clinical/laboratory information being automatically recording in electronic systems. This is unlike questionnaire studies in which missing
data occur due to patient preferences to reveal their details (i.e., misclassification bias).

Fourth, some clinical information may be too sophisticated to be recorded\(^{[39]}\) (e.g., lifestyle factors, dietary pattern, exercises), incompletely or selectively recorded (e.g., smoking, alcohol use, body mass index, family history), or not represented by the coding system (e.g., bowel preparation in colonoscopy research). This may be partially addressed by using other variables as proxies for unmeasured variables. For example, chronic pulmonary obstructive disease is a surrogate marker of heavy smoking. Certainly, in the most ideal situation, adjusting for a perfect proxy of an unmeasured variable achieves the same effect as adjusting for the variable itself. Large healthcare datasets will usually contain a sufficient set of measured surrogate variables, insofar as it represents an overall proxy for relevant unmeasured confounding. A more fascinating and precise approach is the analysis of unstructured data within the electronic health records [e.g., natural language processing (NLP) to extract meaningful data from text-based documents that do not fit into relational tables]\(^{[30]}\). As an example, free-text searches outperformed discharge diagnosis coding in the detection of postoperative complications\(^{[31]}\). In the field of pharmacoepidemiological studies, over-the-counter medication usage is frequently not captured in electronic database systems. These “messy data” (false, imprecise or missing information), more often representing non-differential misclassification bias instead of a differential one, will usually attenuate any positive association, and even trend towards null\(^{[23]}\).

Generally, a “false-negative” result is preferred to a “false-positive” one in epidemiological studies.

Lastly, ethical concerns over an individual’s right to privacy versus the common good have yet to be satisfactorily addressed\(^{[19]}\). The issue of privacy can be tackled with de-identification of individuals using anonymous identifiers (e.g., unique reference keys in terms of numbers and/or letters), although in rare occasions a remote possibility of discerning individuals still exists\(^{[23]}\). For instance, individuals with a very rare disease may be identified via mapping with enough geographical detail.

Although Big Data analysis generates hypothesis-free predictive models wherein no clear explanation accountable for the outcome may be found, it provides a valuable opportunity to derive hypotheses based on these observations, which may not be otherwise conceivable. This strategy (in silico discovery and validation) applies to both candidate biomarkers and therapeutic targets to accelerate the development process for an earlier clinical application. In the end, traditionally hypothesis-driven scientific method research should still be applied to validate the results in multicentre, prospective studies or RCTs. Table 1 summarizes the advantages and shortcomings of Big Data analysis in gastroenterology and hepatology research, as well as its proposed solutions.

**PROPENSITY SCORE METHODOLOGY IN BIG DATA ANALYSIS**

As stated previously, confounding is an inevitable problem of observational studies, irrespective of the sample size. Confounding is a systematic difference between the group with the exposure of interest and the control group\(^{[27]}\). It arises when other factors that affect the exposure of interest are also independent determinants of the outcome. Common sources of confounding include confounding by indication/disease severity, confounding by functional status and cognitive impairment, healthy user/adherer bias, ascertainment bias, surveillance bias, access to healthcare, selective prescription, and the treatment of frail and very sick patients\(^{[27]}\).

Propensity score (PS) methodology has become a widely accepted and popular approach in Big Data analysis of analytic studies in healthcare research. A PS is the propensity (probability) of an individual being assigned to an intervention/exposure conditional on other given covariates, but not the outcome\(^{[33]}\). It is derived from the logistic regression model by regressing the covariates (exclusive of the outcome) onto the exposure of interest. By taking into account this single score in further statistical analysis, a balance of the characteristics between exposure and control groups could theoretically be achieved in the absence of unmeasured confounding. PS methodology entails PS matching, PS stratification/subclassification, PS analysis by inverse probability of treatment weighting, PS regression adjustment, or a combination of these methods, and we refer readers to other articles for further details\(^{[33]}\).

To control for confounding, outcome regression models are traditionally applied. However, this is constrained by the dimensionality of available variables in healthcare datasets (i.e., “curse of dimensionality”). In the simulation study on logistic regression analysis by Peduzzi et al\(^{[34]}\), a low events per variable (EPV) was found to be more
Table 1  Advantages and shortcomings of Big Data analysis (with proposed solutions)

| Advantages                                                                 | Solution                                                                 |
|---------------------------------------------------------------------------|--------------------------------------------------------------------------|
| Clinical data readily available with minimal resources required            | Cross reference with medical records in a subset of the sample           |
| Can study rare exposures                                                  | Statistical methods to deal with missing data, e.g., multiple imputation |
| Can study rare events                                                     | Text mining or natural language processing of unstructured data          |
| Can study long-term effects                                               | Surrogate markers (e.g., COPD for smoking, alcohol-related diseases for  |
| Real-world data                                                           | alcoholism)                                                              |
| Large sample size                                                         | Inclusion of a large set of measured variables                           |
| Subgroup analysis                                                         | Text mining or natural language processing of unstructured data          |
| Sensitivity analysis                                                      | De-identification of individuals                                         |
| Interaction of different variables                                        | Review of study plan by local ethics committee                           |
| Adjustment of outcome to a multitude of risk factors                      | Validation in prospective studies or randomized control trials           |
| Precise estimation of effect size                                         |                                                                         |
| Reliable capture of small variations in incidence or disease flare        |                                                                         |
| No selection bias if \( n = \text{all} \)                                 |                                                                         |

| Shortcomings specific of Big Data analysis                                | Solution                                                                 |
|--------------------------------------------------------------------------|--------------------------------------------------------------------------|
| Data validity                                                            | Cross reference with medical records in a subset of the sample           |
| Missing data                                                             | Statistical methods to deal with missing data, e.g., multiple imputation |
| Incomplete capture of variables or unavailability of certain diagnosis   | Text mining or natural language processing of unstructured data          |
| Privacy                                                                  | Surrogate markers (e.g., COPD for smoking, alcohol-related diseases for  |
| Hypothesis-free predictive models                                         | alcoholism)                                                              |
| Privacy                                                                  | Inclusion of a large set of measured variables                           |
| Privacy                                                                  | Text mining or natural language processing of unstructured data          |
| Privacy                                                                  | De-identification of individuals                                         |
| Hypothesis-free predictive models                                         | Review of study plan by local ethics committee                           |
| Hypothesis-free predictive models                                         | Validation in prospective studies or randomized control trials           |

| Shortcomings of all observational study including Big Data analysis       | Solution                                                                 |
|--------------------------------------------------------------------------|--------------------------------------------------------------------------|
| Residual and/or unmeasured confounding                                    | Cross reference with medical records in a subset of the sample           |
| Reverse causality/protopathic bias (outcome of interest leads to exposure to interest) | Statistical methods to deal with missing data, e.g., multiple imputation |
| Example: Early symptoms of undiagnosed GC leads to PPI use, rather than PPIs cause GC | Text mining or natural language processing of unstructured data          |
| Selection bias                                                           | Surrogate markers (e.g., COPD for smoking, alcohol-related diseases for  |
| Indication bias (or confounding by indication/disease severity)           | alcoholism)                                                              |
| Selection bias                                                           | Inclusion of a large set of measured variables                           |
| Confounding by functional status and cognitive impairment                 | Text mining or natural language processing of unstructured data          |
| Healthy user bias / adherer bias (individuals who are more health conscious tend to have better health outcomes) | De-identification of individuals                                         |
| Example: PPI users may undergo upper endoscopy more frequently than non-PPI users, and hence more GC detected in PPI users | Review of study plan by local ethics committee                           |
| Immortal time bias (arises when the study outcome cannot occur during a period of follow-up due to study design) | Validation in prospective studies or randomized control trials           |
| Ascertaintment bias / surveillance bias / detection bias (differential degree of surveillance or screening for the outcome among exposed and unexposed individuals) | Cross reference with medical records in a subset of the sample           |
| Access to healthcare                                                     | Statistical methods to deal with missing data, e.g., multiple imputation |
| Selective prescription and treatment in frail and very sick patients      | Text mining or natural language processing of unstructured data          |
| COPD: Chronic pulmonary obstructive disease; RCT: Randomized controlled trial; GC: Gastric cancer; PPI: Proton pump inhibitor; PS: Propensity score. |

Cheung KS et al. Big Data in gastrointestinal research
influential than other problems, such as sample size or the total number of events. If the number of EPV is less than ten, the regression coefficients may be biased in both positive and negative directions, the sample variance of the regression coefficients may be over- or under-estimated, the 95% confidence interval may not have proper coverage, and the chance of paradoxical associations (significance in the wrong direction) may be increased. The use of PS methodology, by condensing all covariates into one single variable (PS), can thus address this “curse of dimensionality”\cite{35}. However, PS methodology may not offer additional benefits if the EPV is large enough. Statistical significance differs between the two methods in only 10% of cases, in which traditional regression models give a statistically significant association not otherwise found in PS methodology\cite{36}. In addition, the effect estimate derived by traditional models differs by more than 20% from that obtained by PS methodology in 13% of cases\cite{37}.

The use of PS allows the recognition of subjects with absolute indications (or contraindications) of an intervention, who have no comparable unexposed (or exposed) counterparts for valid estimation of relative or absolute differences in the outcomes\cite{38}. This can be easily identified by plotting a graph of PS distribution between the two groups to look for areas of non-overlap. This pitfall is unlikely to be recognised by traditional modelling, and could be influential as a result of effect measure modification or model misspecification. PS methodology allows trimming (i.e., excluding individuals with areas of non-overlap in PS distributions) or matching to ensure comparability between exposure and control groups. In particular, PS matching does not make strong assumptions of linearity in the relationship of propensity with outcome, and is also better than other matching strategies to achieve an optimal balance of a large set of covariates. The interaction effect of PS with treatment may exist, as effectiveness of an intervention varies according to the indications. An intervention is beneficial in patients with clear indications, but paradoxically provides no benefit, or is even harmful in those with weak indications or contraindications. This was nicely illustrated in the study by Kurth et al\cite{39} on the effect of tissue plasminogen activator on in-hospital mortality. Table 2 summarizes the major advantages of PS methodologies.

**EXAMPLES OF GASTROINTESTINAL DISEASE RESEARCH USING BIG DATA APPROACHES**

Tables 3-7 show a list of research using Big Data approaches from different regions/countries worldwide. This list is by no means exhaustive, however provides a few distinct examples of how Big Data analysis can generate high-quality research outputs in the field of gastroenterology and hepatology. Specifically, in the following section, we will demonstrate how researchers conducted research on some important gastrointestinal and liver diseases, including gastric cancer, gastrointestinal bleeding (GIB), IBD, colorectal cancer (CRC), and HCC. It should be noted that the majority of database systems fulfil the characteristics of the 3Vs (volume, velocity and variety). This is with the exception of the Nurses Health Study (known as NHSII) and Health Professionals Follow-up Study (known as HPFS), which are prospective studies without instantaneous updates of the clinical information using participant questionnaires, thus limiting the velocity of data generation and transformation.

**Gastric cancer**

Gastric cancer is the fifth most common cancer and third leading cause of cancer-related deaths worldwide\cite{39}. Around two-thirds of patients have gastric cancer diagnosed at an advanced stage, rendering curative surgery impossible\cite{40,41}. Infection by *Helicobacter pylori* (*H. pylori*), a class I human carcinogen\cite{42}, confers a two- to three-fold increase in gastric cancer risk\cite{43,44}. RCTs and prospective cohort studies on the effect of *H. pylori* eradication on gastric cancer development are difficult to perform due to the low incidence of gastric cancer, as well as the long lag time of any potential benefits, which mandate a huge sample size with long follow-up duration.

However, Big Data analysis may shed new light on the role of *H. pylori* eradication on gastric cancer development based on population-based health databases. It was shown in a Swedish population-based study that *H. pylori* eradication therapy was associated with a lower gastric cancer risk compared with the general population, but this effect only started to appear beyond 5 years post-treatment\cite{45}. Stratified analysis in a Taiwanese study based on the National Health Insurance Database (commonly known as NHID) showed that early *H. pylori* eradication was associated with a lower gastric cancer risk than late eradication when compared with the general population\cite{46}. Based on a territory-wide public healthcare database in Hong Kong
called the Clinical Data Analysis and Reporting System, *H. pylori* eradication therapy was beneficial even in older age groups (≥ 60 years)\(^\text{[17]}\). Apart from *H. pylori* eradication, regular non-steroidal anti-inflammatory drug use was also shown to be a protective factor for gastric cancer based on the study from NHID from Taiwan\(^\text{[46]}\). Long-term aspirin use further reduced gastric cancer risk in patients who had received *H. pylori* eradication therapy\(^\text{[16]}\). Moreover, the long-term use of metformin was associated with a lower gastric cancer risk in our patients who had received *H. pylori* eradication therapy\(^\text{[16]}\).

On the other hand, long-term proton pump inhibitor (PPI) use was associated with an increased gastric cancer risk in patients who had received *H. pylori* eradication therapy\(^\text{[10]}\), which is otherwise difficult to be addressed by RCTs\(^\text{[8]}\). This finding was echoed by another nationwide study\(^\text{[9]}\). A study on the interaction between aspirin and PPIs further showed that PPIs were associated with a higher cancer risk among non-aspirin users, but not among aspirin users\(^\text{[44]}\). However, pantoprazole, a long-acting PPI, was not associated with an increased gastric cancer risk compared with other shorter-acting PPIs in a United States Food and Drug Administration (commonly known as FDA)-mandated study\(^\text{[9]}\). Other risk factors for gastric cancer determined by large healthcare datasets included the extent of gastric intestinal metaplasia, as well as a family history of gastric cancer\(^\text{[12]}\). In addition, racial/ethnic minorities had a 40%-50% increase in gastric cancer risk compared with the Hispanic and white populations\(^\text{[11]}\).

**GIB**

Upper GIB is one of the most common causes of hospitalization, and emergency department visits that pose significant economic burdens on the healthcare system. Antiplatelet agents (including aspirin and P2Y\(_{12}\) inhibitors) were major causative agents\(^\text{[11]}\). In a nationwide retrospective cohort study, it was shown that *H. pylori* eradication and PPIs were associated with reduced incidences of gastric ulcer (42%-48%) and duodenal ulcers (41%-71%)\(^\text{[11]}\). However, importantly, concomitant use of clopidogrel, H2-receptor antagonists (referred to as H2RAs) and PPIs was associated with an increased risk of acute coronary syndrome or all-cause mortality\(^\text{[12]}\). This harmful effect was particularly prominent for PPIs with high CYP2C19 inhibitory potential\(^\text{[12]}\). These findings raised the need for judicious use of gastroprotective agents in clopidogrel users, and called for further studies to determine causality versus biases (e.g., indication bias).

When novel oral anticoagulants (NOACs) were first introduced, there was a paucity of real-world data on the GIB risk and its preventive measures\(^\text{[11]}\). In a territory-wide retrospective cohort study, the risk of GIB was determined in dabigatran users, with risk factors identified and effects of gastroprotective agents (PPIs and H2RAs) investigated\(^\text{[11]}\). All patients who were newly prescribed dabigatran were identified (n = 5041). There were 124 (2.5%) GIB cases, with an incidence rate of GIB of 41.7 cases per 1,000 person-years. PPIs were found to protect against upper GIB. This important finding has recently been echoed by an even larger-scale study\(^\text{[63]}\), with a consistent beneficial effect of PPIs on upper GIB across various NOACs (dabigatran, rivaroxaban and apixaban). Head-to-head comparisons between different NOACs and their interaction with PPIs would barely be possible in other study designs, given the huge number of study subjects required to ensure statistical power. These drug safety data can be easily ascertained by Big Data analysis of electronic health databases, which would be otherwise difficult in other observational studies or RCTs due to the various limitations previously mentioned, especially if the absolute risk difference is small.

**IBD**

Precise outcome prediction in IBD remains challenging, as it is a highly heterogeneous...
## Table 3  Examples of studies on gastric cancer research by utilization of large healthcare datasets

| Country/Region | Database | Area of research | Sample size | Design, statistical methods and 3V | Application |
|----------------|----------|------------------|-------------|-----------------------------------|-------------|
| Taiwan, China  | Taiwan National Health Insurance Database (NHID) | GC | 80255 | Nationwide retrospective cohort study | Early vs late *H. pylori* eradication on GC risk |
|                | Wu et al[46], 2009 | GC | 52161 | Nationwide retrospective cohort study | Association between NSAIDs and GC |
|                | Wu et al[46], 2010 | GC | 52161 | Nationwide retrospective cohort study | Association between NSAIDs and GC |
| Hong Kong, China | Clinical Data Analysis and Reporting System (CDARS) | GC | 63397 | Territory-wide retrospective cohort study | Association between PPIs and GC |
|                | Cheung et al[49], 2018 | GC | 63605 | Territory-wide retrospective cohort study | Association between aspirin and GC |
|                | Cheung et al[49], 2018 | GC | 63397 | Territory-wide retrospective cohort study | Effect of *H. pylori* eradication among different age groups |
|                | Cheung et al[49], 2018 | GC | 7266 | Territory-wide retrospective cohort study | Association between metformin and GC |
| Sweden         | Swedish Cancer Registry | GC | 797067 | Nationwide retrospective cohort study | Association between PPIs and GC |
|                | Swedish Prescribed Drug Registry | GC | 95176 | Nationwide retrospective cohort study | Effect of *H. pylori* eradication on GC risk |
|                | Brusselaers et al[52], 2017 | GC | 95176 | Nationwide retrospective cohort study | Effect of *H. pylori* eradication on GC risk |
|                | Doorakkers et al[52], 2018 | GC | 95176 | Nationwide retrospective cohort study | Effect of *H. pylori* eradication on GC risk |
disease with numerous predictive factors. Machine learning algorithms are particularly useful in deriving predictive models, including risk factors, disease outcomes, and treatment responses, hence allowing the identification of at-risk individuals who require early aggressive intervention. Today, there is still an unmet need for newer therapeutic agents for IBD, as the long-term efficacy of current options including anti-tumour necrosis factor (anti-TNF) and anti-integrin αβ7, are still unsatisfactory. However, the process of new drug discovery for IBD is prolonged and costly, and success is not guaranteed. For instance, mongersen, an anti-IL-17A monoclonal antibody, was also disappointing in moderate to severe Crohn’s disease, in which it was less effective and carried higher rates of adverse events compared with placebo, despite the potential role of IL-17 in Crohn’s disease as suggested by animal models and GWAS. Drug repurposing from Big Data applications helps in this regard, as illustrated by Dudley et al. In that study, computational approaches were used to discover new drugs for IBD in silico by comparing the gene expression profiles from 164 drug compounds to a gene expression signature of IBD from publicly available data obtained from the NCBI Gene Expression Omnibus. A technique, called “signature inversion”, was used to identify drugs that can reverse a disease signature (transcriptomic, proteomic, or other surrogate markers of disease activity). Topiramate, an FDA-approved drug for treating epilepsy, was identified to be a potential therapeutic drug in IBD with experimental validation in a mouse model. The potential role of topiramate, however, was later refuted by a retrospective cohort study, and no further studies have been conducted.

As discussed previously, some diseases may not be coded in the electronic database. As an example, the effects of anti-TNF versus vedolizumab on arthralgia in IBD patients were studied using NLP. As the electronic coding of arthralgia is not commonly performed in gastroenterology practices, Cai et al. used NLP to directly extract this non-structured information from the narrative electronic medical records, and converted it into a structured variable (joint pain: yes/no) of analysis. Without NLP, simply relying on a diagnosis code may bias any potential positive association towards null. On the other hand, manual review of the electronic medical records demands an intensive input of manpower, and accuracy is also not fully guaranteed.

In a study that involved 827,239 children, antibiotics exposure during pregnancy was found to be associated with an increased risk of very early onset IBD. This study was achieved by merging data from several databases with the unique personal identity number assigned to Swedish residents. One of the databases, the Swedish Medical Birth Register, enabled the identification of child-mother links. This study illustrates the unique role of Big Data applications in investigating childhood exposure that affects disease development in adulthood, which is nearly impossible in the setting of RCT (ethical and resource issue) and other types of observational studies (e.g., recall bias, resource issue).

CRC
CRC is the third most common cancer and the second leading cause of cancer-related death. As a period of 10 years is required for the development of the adenoma-carcinoma sequence, identification of risk factors of CRC would have been difficult with RCTs. A large number of high-quality research has been conducted based on the NHS, NHSSII and HPFS cohorts. Type II diabetes mellitus was associated with a 1.4-fold increase in CRC risk. A positive association between obesity and early-onset CRC also existed among women. Some of the risk factors (e.g., smoking, body mass index, alcohol intake) and protective factors (e.g., physical activity, folate and calcium intake) of CRC were found to be associated with the development of its precursors, adenomas and/or serrated polyps. Among non-metastatic CRC patients, higher
Table 4  Examples of studies on gastrointestinal bleeding and/or proton pump inhibitor research by utilization of large healthcare datasets

| Country/Region          | Database                                           | Area of research | Sample size | Design, statistical methods and 3V                  | Application                                      |
|------------------------|----------------------------------------------------|------------------|-------------|----------------------------------------------------|--------------------------------------------------|
| Taiwan, China          | Taiwan National Health Insurance Database (NHID)   | PUD              | 403567      | Nationwide retrospective cohort study              | Effect of H. pylori therapy and PPIs on PUD      |
|                        |                                                   |                  |             | Volume, Velocity and Variety                       |                                                  |
|                        |                                                   | PUD              | 32235       | Nationwide retrospective cohort study              | Risk of rebleeding from PUD in ESRD patients     |
|                        |                                                   |                  |             | Volume, Velocity and Variety                       |                                                  |
|                        |                                                   | PPIs             | 6552        | Nationwide retrospective cohort study              | Effect of clopidogrel and PPIs on ACS            |
|                        |                                                   |                  |             | Volume, Velocity and Variety                       |                                                  |
| South Korea            | Korean Health Insurance Review and Assessment Service (HIRA) | PPIs            | 92233       | Nationwide retrospective cohort study              | Effect of PPIs on thrombotic risk                |
|                        |                                                   |                  |             | Volume, Velocity and Variety                       |                                                  |
| Hong Kong, China       | Clinical Data Analysis and Reporting System (CDARS) | Dabigatran       | 5041        | Territory-wide retrospective cohort study          | Risk factors for dabigatran-associated gastrointestinal bleeding |
|                        |                                                   |                  |             | Volume, Velocity and Variety                       |                                                  |

This list is not exhaustive, but serves to provide a few distinct examples of how Big Data analysis can generate high-quality research outputs in the field of gastroenterology and hepatology. 3V: Volume/velocity/variety; PUD: Peptic ulcer disease; H. pylori: Helicobacter pylori; PPIs: Proton pump inhibitors; ESRD: End-stage renal disease; ACS: Acute coronary syndrome.

coffee\(^{[77]}\), calcium\(^{[78]}\) and fibre\(^{[79]}\) intake were found to be associated with a lower CRC-specific and all-cause mortality. Concerning hereditary cancer syndromes, the Dutch Lynch syndrome Registry is one eminent example of the hereditary cancer registries. It was noted that surveillance could reduce CRC-related mortality\(^{[80]}\). However, in a subsequent study involving three countries (the Netherlands, Germany and Finland) with different surveillance policies, a shorter surveillance colonoscopy interval (annually) was not associated with a reduction in CRC when compared with longer intervals (1-2 yearly and 2-3 yearly intervals)\(^{[81]}\). The Dutch polyposis registry is another example that includes adenomatous polyposis coli patients\(^{[82]}\).

**HCC**

Chronic hepatitis B virus (HBV) infection is a major public health threat that results in significant morbidity and mortality\(^{[83]}\). The prevalence of chronic HBV infection was estimated at 3.5% (257 million people) worldwide in 2016. Major complications of chronic HBV infection included HBV reactivation with hepatitis flare\(^{[84]}\), cirrhosis and HCC\(^{[85,86]}\).

Nucleos(t)ide analogue (NA) therapy was found to be associated with a lower HCC risk among chronic hepatitis B (CHB) patients\(^{[87]}\). This was in line with the finding from an ecologic study showing that NA therapy was associated with a reduction in age-adjusted liver cancer incidence\(^{[88]}\). The beneficial effect of NA was further proven among CHB patients who had undergone liver resection for HCC, in which NA therapy was associated with a lower risk of HCC recurrence\(^{[89]}\). The recent finding that tenofovir was associated with around a 40% reduction in HCC risk compared with entecavir has guided the choice of antiviral therapy in CHB patients at high risk of HCC (e.g., cirrhosis)\(^{[90]}\). Although diabetes mellitus was associated with an increased HCC risk\(^{[91]}\), each incremental year increase in metformin use resulted in a 7% reduction in HCC risk for diabetic patients.

The choices of therapeutics drugs for HCC are still currently limited. Big Data
Table 5  Examples of studies on inflammatory bowel disease research by utilization of large healthcare datasets

| Country/Region | Database | Area of research | Sample size | Design, statistical methods and 3V | Application |
|----------------|----------|------------------|-------------|-----------------------------------|-------------|
| South Korea    | Korean Health Insurance Review and Assessment Service (HIRA) | UC | 11233 | Nationwide retrospective cohort study, Comparator: general population, Volume, Velocity and Variety | Incidence and clinical impact of perianal disease in UC |
| Taiwan, China  | Taiwan National Health Insurance Database (NHID) | IBD | 38039 | Nationwide retrospective cohort study to compare IBD patients with general population to derive SIR | Association between IBD and herpes zoster infection |
| Sweden         | Swedish Patient Registry | UC | 63711 | Nationwide retrospective cohort study | Association between appendectomy and UC |
|                | Swedish Medical Birth Register (child-mother link) | IBD | 827,239 children born between 2006 and 2013 | Nationwide prospective population-based register study | Association between maternal exposure to antibiotics during pregnancy and very early onset IBD in adulthood |
|                | Swedish Multigeneration Register (child-father link) | IBD | 827,239 children born between 2006 and 2013 | Nationwide prospective population-based register study | Association between maternal exposure to antibiotics during pregnancy and very early onset IBD in adulthood |
|                | Swedish Prescribed Drug Register National Patient Register | IBD | 827,239 children born between 2006 and 2013 | Nationwide prospective population-based register study | Association between maternal exposure to antibiotics during pregnancy and very early onset IBD in adulthood |
| United States  | NCBI Gene Expression Omnibus (GEO) | IBD | n.a. | Signature inversion study | Topiramate as a potential therapeutic agent against IBD |
|                | n.a. | IBD | 1585 | Retrospective cohort study Natural language processing | Association between arthralgia and biologics (anti-TNF vs vedolizumab) |
| n.a.           | International IBD Genetics Consortium's Immunochip project | IBD | 53279 | Machine learning algorithm | Predictors of IBD |
| United States  | n.a. | IBD | 573 colonoscopy reports | Retrospective cohort study Natural language processing | Differentiation of surveillance from non-surveillance colonoscopy |
|                | Hou et al[80], 2013 | IBD | 1080 | Retrospective cohort study Random Forest machine learning algorithm | Prediction of IBD remission in thiopurine users |
|                | Waljee et al[81], 2017 | IBD | 20368 | Retrospective cohort study Random Forest machine learning algorithm | Prediction of hospitalization and outpatient steroid use |
|                | Waljee et al[81], 2017 | IBD | 491 | Retrospective cohort study Random Forest machine learning algorithm | Prediction of steroid-free endoscopic remission with vedolizumab in UC |
approaches in drug repurposing have once again shed light on the potential anti-cancer role of some medications currently approved for other purposes. For example, Chen et al. collected publicly available data from HCC studies on HCC-related genes, and 6,100 drug-mediated expression profiles from Connectivity Map, which is a search engine cataloguing the effects of pharmacological compounds on different cell types. By using “signature inversion” approaches, chlorpromazine and trifluoperazine were found to have anti-cancer effects on HCC. Another study using a similar computational approach unveiled the potential anti-HCC effect of prenylamine.

FUTURE PERSPECTIVE OF BIG DATA RESEARCH

Clinicians and scientists in the field of gastroenterology and hepatology should aspire to optimize the potential advantage of powerful Big Data in translating routine clinically-collected data into precision medicine, the development of new biomarkers, and therapeutic agents in a relatively short and effective manner for preventing diseases and/or improving patient outcomes. However, some areas are still primitive or under-explored.

Parent-child linkage is one of the examples unique to Big Data analysis. Parental factors could have important bearings on the development of various diseases during childhood. One example is linking racial/ethnic and socioeconomic data from both parents with childhood obesity. As for gastrointestinal and liver diseases, one study showed that maternal use of antibiotics during pregnancy was associated with an increased risk of very early onset IBD. One possible mechanism is via the alteration of the gut microbiome. However, the unavailability of direct linkage is still a major issue that can only be partly addressed by indirect inference, such as a probabilistic linkage of maternal and baby healthcare characteristics. It is therefore imperative to have a database system that has direct parent-child linkages, of which many of the currently existing electronic databases are still devoid.

Drug safety is another field that could benefit from Big Data research. First, preclinical computational exclusion of potentially toxic drugs will improve patient safety while reducing the delay in drug discovery and expense. Second, the efficiency of post-marketing surveillance on drug toxicities can be enhanced. Concerning the missing data for some important risk factors (e.g., smoking, alcohol intake, body mass index), administering institutions should be aware of the immense potential of Big Data, and take pre-emptive actions to start collecting these data. Although the hypothesis-free approach of Big Data analysis facilitates the discovery of new biomarkers and drugs, the results should still be validated in multi-centres. A network involving multiple centres across nations should be established to foster a centralized, comprehensive collection and validation of data. While patient privacy should be upheld, regulatory mechanisms should be realistically enforced without jeopardizing the conduct of Big Data research.

CONCLUSION

The advent of Big Data analysis in medical research has revolutionized the traditional hypothesis-driven approach. Big Data analysis provides an invaluable opportunity to improve individual and public health. Data fusion of different sources will enable the analysis of health data from different perspectives across different regions. In this era of digitalized healthcare research and resources, manpower and time are no longer hurdles to the production of high-quality clinical studies in a cost-effective manner. With continuous technological advancements, some of the current limitations with Big Data may be further minimized.
## Table 6: Examples of studies on colorectal cancer research by utilization of large healthcare datasets

| Country/Region           | Database                                                                 | Area of research | Sample size | Design, statistical methods and 3V Application |
|--------------------------|--------------------------------------------------------------------------|------------------|-------------|-----------------------------------------------|
| Hong Kong, China         | Clinical Data Analysis and Reporting System (CDARS)                       | CRC Cheung et al[101], 2019 | 197902      | Territory-wide retrospective cohort study      |
|                          |                                                                          | CRC              | 187897      | Volume, Velocity and Variety                  |
|                          |                                                                          |                  |             | Epidemiology, characteristics, risk factors and prognosis of postcolonoscopy Colorectal cancer in Asians |
| United States            | Nurses’ Health Study II (NHSII)                                          | CRC Ma et al[69], 2018 | 134763      | Territory-wide prospective cohort study       |
|                          | Nurses’ Health Study (NHS)                                               | CRC Yang et al[69], 2018 | 1660        | Association between statins and CRC          |
|                          | Health Professionals Follow-up Study (HPFS)                              | CRC Hu et al[69], 2018 | 1599        | Association between DM and CRC                |
|                          |                                                                          | CRC Song et al[69], 2018 | 1575        | Effect of calcium intake, coffee and fibre on survival after CRC diagnosis |
|                          | Nurses’ Health Study (NHS)                                               | CRC He et al[69], 2006 | 141143      | Risk factors of serrated polyps and conventional adenomas |
|                          | Nurses’ Health Study II (NHSII)                                          | CRC              | 85256       | Association between obesity and CRC           |
|                          | Health Professionals Follow-up Study (HPFS)                              |                  |             | Volume and Variety                            |
| Netherlands              | Dutch Lynch syndrome Registry                                            | Various cancers including CRC | 2788        | Decrease in CRC-related mortality in Lynch syndrome families by surveillance |
| Netherlands, Germany, Finland | Dutch Lynch syndrome Registry                                      | CRC Engel et al[11], 2018 | 2747 patients with 16327 colonoscopies | Retrospective cohort study                  |
|                          | German HPNCC Consortium Finland                                          |                  |             | Volume, Velocity and Variety                  |
|                          |                                                                          |                  |             | Surveillance interval on CRC incidence and stage |

This list is not exhaustive, but serves to provide a few distinct examples of how Big Data analysis can generate high-quality research outputs in the field of gastroenterology and hepatology. 3V: Volume/velocity/variety; CRC: Colorectal cancer; DM: Diabetes mellitus.

## Table 7: Examples of studies on hepatocellular carcinoma research by utilization of large healthcare datasets

| Country/Region           | Database                                                                 | Area of research | Sample size | Design, statistical methods and 3V Application |
|--------------------------|--------------------------------------------------------------------------|------------------|-------------|-----------------------------------------------|
| Taiwan, China            | Publicly available data on HCC-related genes Connectivity Map (CMap) -- includes 6100 drug-mediated expression profiles | HCC Chen et al[17], 2011 | n.a.        | Signature inversion study                     |
|                          |                                                                          |                  |             | Volume, Velocity and Variety                  |
|                          |                                                                          |                  |             | Anti-cancer effects of chlorpromazine and trifluoperazine on HCC |
This list is not exhaustive, but serves to provide a few distinct examples of how Big Data analysis can generate high-quality research outputs in the field of gastroenterology and hepatology. 3V: Volume/velocity/variety; HCC: Hepatocellular carcinoma; NA: Nucleos(t)ide analogue; DM: Diabetes mellitus; PS: Propensity score; CHB: Chronic hepatitis B; SIR: Standardized incidence ratio; HDV: Hepatitis D virus.

| Country | Database/Linux | Dataset | Study Type | Study Details | Research Outputs |
|---------|----------------|---------|------------|---------------|------------------|
| Sweden  | Swedish Cancer Registry | HCC | Nationwide retrospective cohort study | Nationwide retrospective cohort study with a comparison of 9,160 CHB patients | Association between concomitant HBV/HDV infection and HCC |
|         | Swedish Patient Registry | HCC | Nationwide retrospective cohort study | Nationwide retrospective cohort study with a comparison of 9,160 CHB patients | Association between concomitant HBV/HDV infection and HCC |
|         | Taiwan National Health Insurance Database (NHID) | HCC Wa et al[8], 2012 | Nationwide retrospective cohort study | Nationwide retrospective cohort study with a comparison of 4,569 CHB patients | Association between NA therapy and HCC recurrence among patients with HBV-related HCC after liver resection |
|         | Taiwan National Health Insurance Database (NHID) | HCC Chen et al[9], 2013 | Nationwide case-control study | Nationwide case-control study with a comparison of 29,290 CHB patients | Association between DM and HCC |
|         | Taiwan National Health Insurance Database (NHID) | HCC Wu et al[8], 2014 | Nationwide retrospective cohort study | Nationwide retrospective cohort study with a comparison of 43,190 CHB patients | Association between NA therapy and HCC among CHB patients |

Cheung KS et al. Big Data in gastrointestinal research

REFERENCES

1. Lohr S. The Origins of ‘Big Data’: An Etymological Detective Story [cited 25 January 2019]. Available from: https://blogs.nytimes.com/2013/02/01/the-origins-of-big-data-an-etymological-detective-story
2. Kitchin R, McArthur G. What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. Big Data Soc 2016; 1:103 [DOI: 10.1177/2053951716631130]
3. Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Byers AH. Big Data: The Next Frontier for Innovation, Competition, and Productivity [cited 25 January 2019]. Available from: https://bigdatawg.nist.gov/pdf/MG_big_data_full_report.pdf
4. Nickerson DW, Rogers T. Political campaigns and big data. J Econom Perspect 2014; 28: 51-74 [DOI: 10.1257/jep.28.2.51]
5. Laney D. 3D data management: controlling data volume, velocity and variety [cited 25 January 2019]. Available from: https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf
6. European Commission. Study on Big Data in Public Health, Telemedicine and Healthcare. December 2016 [DOI: 10.2875/734795]
7. Alonso SG, de la Torre Diez I, Rodrigues JPC, Hamrioui S, López-Corona M. A Systematic Review of Techniques and Sources of Big Data in the Healthcare Sector. J Med Syst 2017; 41: 183 [PMID: 29032458 DOI: 10.1007/s10916-017-0832-2]
8. Bellazzi R. Big data and biomedical informatics: A challenging opportunity. Yearb Med Inform 2014; 9: 8-13 [PMID: 24853034 DOI: 10.15265/iy.2014-0024]
9. Ollivera P, Danese S, Jay N, Natoli G, Peyrin-Biroulet L. Big data in IBD: A look into the future. Nat Rev Gastroenterol Hepatol 2019; 16: 224-234 [PMID: 30659247 DOI: 10.1038/s41575-019-0102-5]
10. Mirkov MV, Verstockt B, Cleynen I. Genetics of inflammatory bowel disease: Beyond NOD2. Lancet Gastroenterol Hepatol 2017; 2: 224-234 [PMID: 28404137 DOI: 10.1016/s2468-1253(16)30111-x]
11. Shivade C, Raghavan P, Posler-Lussier E, Embi PJ, Elhadad N, Johnson SB, Lai AM. A review of approaches to identifying patient phenotype cohorts using electronic health records. J Am Med Inform
Global Burden of Disease Cancer Collaboration. Fitzmaurice C, Allen C, Barber RM, Barregard L, et al. Cancer conditions of nonuniform effect. Am J Epidemiol 2006; 164: 437-447 [PMID: 16632131 DOI: 10.1016/j.jclinepi.2005.07.004]

Kurth T, Walker AM, Glynn RJ, Chan KA, Gaziano JM, Berger K, Robins JM. Results of multivariable propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. J Clin Epidemiol 2005; 58: 550-559 [PMID: 15878468 DOI: 10.1016/S0895-4356(04)00236-3]

Vayena E, Salathé M, Madoff LC, Brownstein JS. Ethical challenges of big data in public health. PLoS Comput Biol 2015; 11: e1004390 DOI: 10.1371/journal.pcbi.1004390

Brookhart MA, Stürmer T, Glynn RJ, Rassen J, Schneeweiss S. Confounding control in healthcare database research: Challenges and potential approaches. Med Care 2010; 48: S114-S120 DOI: 10.1097/MLR.0b013e3181bee3b3

White IR, Royston P. Imputing missing covariate values for the Cox model. Stat Med 2009; 28: 1982-1998 [PMID: 19452569 DOI: 10.1002/sim.3618]

Murdoch TB, Detsky AS. The inevitable application of big data to healthcare care. JAMA 2013; 309: 1351-1352 [PMID: 23549579 DOI: 10.1001/jama.2013.393]

Murff HJ, FitzHenry F, Matheny ME, Gentry N, Kotter KL, Crimmin K, Dittus RS, Rosen AK, Elklin PL, Brown SH, Speroff T. Automated identification of postoperative complications within an electronic medical record using natural language processing. JAMA 2011; 306: 848-855 [PMID: 21862746 DOI: 10.1001/jama.2011.1204]

Cheung KS et al. Big Data in gastrointestinal research

WJG | https://www.wjgnet.com

June 28, 2019 | Volume 25 | Issue 24 |
Cheung KS et al. Big Data in gastrointestinal research

Bhutta ZA, Brenner H, Dicker DJ, Chimed-Oorchir O, Dandona R, Dandona L, Fleming T, Forouzanfar MH, Hancock J, Hay RJ, Hunter-Merrill R, Huynh C, Hosgood HD, Johnson CO, Jonas JB, Khubchandani J, Kumar GA, Kutn M, Lan Q, Larson HJ, Liang X, Lim SS, Lopez AD, McIntyre MF, Marczuk L, Marquez N, Meldad AH, Pinho C, Pourmalek F, Salomon JA, Sanabria JR, Sandar L, Sartorius B, Schwartz SM, Shackelford KA, Shibuya K, Stanaway J, Steiner C, Sun J, Takahashi K, Vollset SE, Vos T, Wagner JA, Wang H, Westerman R, Zeeb H, Zeeck L, Abd-Allah F, Ahmed MB, Alabed S, Alam NK, Alahdher SF, Alem G, Alemayehou MA, Ali R, Al-Raddadi R, Amare A, Amono Y, Artaman A, Asayesh H, Atafai N, Awwadli A, Saleemi HB, Barac A, Bedi N, Benerooi I, Berhanhe A, Bernabé E, Betsu B, Binninga A, Boneya D, Campos-Nonoato I, Castañeda C, Chifflee A, Choi JY, Cowie B, Damtew S, das Neves J, Dey S, Dharmaratne S, Dhillon P, Ding E, Driscoll T, Ekwueme D, Endries AY, Farvid M, Farzadfar F, Fernandes J, Fischer F, G/Hiwot TT, Gebru A, Gopalan S, Hailu A, Horimo M, Horita N, Hussein E, Huybrechts I, Inoue M, Islam F, Jakovljevic M, James S, Javanbakht M, Jose SH, Kasaee A, Kedir MS, Kladar YS, Khang YH, Kim D, Leigh J, Lima D, Malat DC, El Razek HMA, Malekdek R, Malata DC, Marcenes W, Markos D, Melaku YA, Meles KG, Mendoza W, Mengiste DT, Mereteja TJ, Miller TR, Mohammad KA, Mohammad A, Mohammed S, Moradi-Lakeh M, Nagel G, Nand D, Le Nguyen Q, Nolte S, Ogbo FA, Oladimeji KE, Oren E, Park M, Pereira DM, Plass D, Qorbani M, Radfar A, Rafay A, Rahman M, Rana SM, Soreide K, Sapitathy M, Sawhney M, Sepanlou SG, Shaiikh MA, She J, Shiue I, Shore HR, Shrive MG, So S, Soneji S, Stathopoulos V, Stroumpoulis K, Subiyam MB, Sykes BL, Tabarés-Seisdedos R, Tadese F, Teiwa BD, Tessema GA, Thakur JS, Tran BX, Ukwaja KN, Uzochukwu BSC, Vlassov VV, Wiederpass E, Wubsheft Terefe E, Yehyo HG, Yimam HH, Yoninou M, Yu C, Zaïdi Z, Zakri MES, Zenebe ZM, Murray CJL, Naghavi M. Global, Regional, and National Cancer Incidence, Mortality, Years of Life Lost, Years Lived with Disability, and Disability-Adjusted Life-years for 32 Cancer Groups, 1990 to 2015: A Systematic Analysis for the Global Burden of Disease Study. *JAMA Oncol* 2017; 3: 524-548 DOI: 10.1001/jamaoncol.2016.5688

Cervantes A, Roda D, Tarazona N, Roselló S, Pérez-Fidalgo JA. Current questions for the treatment of advanced gastric cancer. *Cancer Treat Rev* 2013; 39: 60-67 DOI: 23102520 DOI: 10.1016/j.cerr.2012.09.007

Van Cutsem E, Sagera T, Topal B, Haustermans K, Prenen H. Gastric cancer. *Lancet* 2016; 388: 2654-2664 DOI: 27156933 DOI: 10.1016/s0140-6736(16)30354-3

Infection with Helicobacter pylori. *IARC Monogr Eval Carcinog Risks Hum* 1994; 61: 177-240 DOI: 7156970

Cavaleiro-Pinto M, Peleteiro B, Lunet N, Barros H. Helicobacter pylori infection and gastric cancer: Systematic review and meta-analysis. *Cancer Causes Control* 2011; 22: 375-387 DOI: 2184266 DOI: 10.1007/s10552-010-9707-2

Cheung KS, Leung WK. Risk of gastric cancer development after eradication of Helicobacter pylori. *World J Gastrointest Oncol* 2018; 10: 115-123 DOI: 2977017 DOI: 10.4251/wjgo.v10.i5.115

Doorakkers E, Lagergren J, Engstrand L, Brusselaars N. Helicobacter pylori eradication treatment and the risk of gastric adenocarcinoma in a Western population. *Gut* 2018; 67: 2092-2096 DOI: 29382776 DOI: 10.1136/gutjnl-2017-315363

Wu CY, Kuo KN, Wu MS, Chen YJ, Lin JT, Effective reduction of gastric cancer risk with proton pump inhibitors: A nationwide population-based cohort study in Sweden. *Br J Cancer* 2017; 28: 28-35 DOI: 29089382 DOI: 10.1136/bmjopen-2017-017739

Leung WK, Wong IOL, Cheung KS, Yeung KF, Chan EW, Wong AYS, Chen L, Wong ICK, Graham DY. Effects of Helicobacter pylori Treatment on Incidence of Gastric Cancer in Older Individuals. *Gastroenterology* 2018; 155: 67-75 DOI: 29505952 DOI: 10.1053/j.gastro.2018.03.023

Wu CY, Wu MD, Kuo KN, Chan EW, Wong CB, Lin JT. Effect of non-steroidal anti-inflammatory drugs on gastric cancer risk: a population-based study. *Gut* 2018; 67: 28-35 DOI: 29089382 DOI: 10.1136/gutjnl-2017-314605

Cheung KS, Chan EW, Wong AYS, Chen L, Seto WK, Leung WK. Aspirin and Risk of Gastric Cancer After Helicobacter pylori Eradication: A Territory-Wide Study. *J Clin Oncol* 2010; 28: 2952-2957 DOI: 20479409 DOI: 10.1002/jco.2009.26.0695

Cheung KS, Chan EW, Chan L, Leung WK. Modifying gastric cancer risk associated with proton pump inhibitors by aspirin after Helicobacter pylori eradication. *Oncotarget* 2018; 9: 36891-36893 DOI: 30229127 DOI: 10.18632/oncotarget.26382

Schneider JL, Kolitsopoulos F, Corley DA. Risk of gastric cancer, gastrointestinal cancers and other cancers: A comparison of treatment with pantoprazole and other proton pump inhibitors. *Aliment Pharmacol Ther* 2016; 43: 73-82 DOI: 26541663 DOI: 10.1111/apt.13458

Reddy KM, Chang IJ, Shi JM, Wu BU. Risk of Gastric Cancer Among Patients With Intestinal Metaplasia of the Stomach in a US Integrated Health Care System. *Clin Gastroenterol Hepatol* 2016; 14: 1420-1425 DOI: 27317852 DOI: 10.1016/j.cgh.2016.05.045

Dong E, Duan L, Wu BU. Racial and Ethnic Minorities at Increased Risk for Gastric Cancer in a Regional US Population Study. *Clin Gastroenterol Hepatol* 2017; 15: 511-517 DOI: 27936564 DOI: 10.1016/j.cgh.2016.11.033

Wu CY, Wu CH, Wu MS, Wang CB, Cheng JS, Kuo KN, Lin JT. A nationwide population-based cohort study shows reduced hospitalization for peptic ulcer disease associated with H pylori eradication and proton pump inhibitor use. *Clin Gastroenterol Hepatol* 2009; 7: 427-431 DOI: 19264578 DOI: 10.1016/j.cgh.2008.12.029

Wu CY, Chan FK, Wu MS, Kuo KN, Wang CB, Tsao CR, Lin JT. Histamin2-receptor antagonists are an...
alternative to proton pump inhibitor in patients receiving clopidogrel. *Gastroenterology* 2010; 139: 1165-1171 [PMID: 20600012 DOI: 10.1053/j.gastro.2010.06.067]

60 Lee TY, Lin JT, Zeng YS, Chen YJ, Wu MS, Wu CY. Association between nucleos(t)ide analog and tumor recurrence in hepatitis B virus-related hepatocellular carcinoma after radiofrequency ablation. *Hepatology* 2016; 63: 1517-1527 [PMID: 26426978 DOI: 10.1002/hep.28266]

61 Cheung KS, Leung WK. Gastrointestinal bleeding in patients on novel oral anticoagulants: Risk, prevention and management. *World J Gastroenterol* 2017; 23: 1954-1963 [PMID: 28373761 DOI: 10.3748/wjg.v23.i11.1954]

62 Chan EW, Lau WC, Leung WK, Mok MT, He Y, Tong TS, Wong JC. Prevention of Dabigatran-Related Gastrointestinal Bleeding With Gastroprotective Agents: A Population-Based Study. *Gastroenterology* 2015; 149: 586-95.e3 [PMID: 25960019 DOI: 10.1053/j.gastro.2015.05.002]

63 Ray WA, Chung CP, Murray KT, Smalley WE, Daughtery JR, Dupont WD, Stein CM. Association of Oral Anticoagulants and Proton Pump Inhibitor Cotherapy With Hospitalization for Upper Gastrointestinal Tract Bleeding. *JAMA* 2018; 320: 2221-2230 [PMID: 30512099 DOI: 10.1001/jama.2018.17242]

64 Wei Z, Wang W, Bradfield J, Li J, Cardinale C, Frackelton E, Kim C, Menich F, Van Steen K, Visscher PM, Baldassano RN, Hakonarson H. International IBD Genetics Consortium. Large sample size, wide variant spectrum, and advanced machine-learning technique boost risk prediction for inflammatory bowel disease. *Am J Hum Genet* 2013; 92: 1008-1012 [PMID: 23731543 DOI: 10.1016/j.ajhg.2013.05.002]

65 Waljee AK, Lipson R, Wittal WL, Yang Z, Liu B, Zhu J, Wallace B, Gavoni SM, Studham RW, Hayward R, Higgins PDR. Predicting Hospitalization and Outpatient Corticosteroid Use in Inflammatory Bowel Disease Patients Using Machine Learning. *Inflamm Bowel Dis* 2017; 24: 45-53 [PMID: 29272474 DOI: 10.1093/ibd/iex007]

66 Waljee AK, Sauder K, Patel A, Segar S, Liu B, Zhang Y, Zhu J, Studham RW, Balis U, Higgins PDR. Machine Learning Algorithms for Objective Remission and Clinical Outcomes with Thiopurines. *J Crohns Colitis* 2017; 11: 801-810 [PMID: 2833183 DOI: 10.1093/jcc/jjx014]

67 Waljee AK, Liu B, Sauder K, Zhu J, Gavoni SM, Studham RW, Higgins PDR. Predicting corticosteroid-free endoscopic remission with vedolizumab in ulcerative colitis. *Aliment Pharmacol Ther* 2018; 47: 763-772 [PMID: 29359519 DOI: 10.1111/apt.14510]

68 Celgene. Celgene provides update on GED-0301 (mongersen) inflammatory bowel disease program [cited 25 January 2019]. Available from: https://ir.celgene.com/press-releases/press-release-details/2017/Celgene-Provides-Update-on-GED-0301-mongersen-Inflammatory-Bowel-Disease-Program/default.aspx

69 Cheung KS, Chen L, Chan EW, Seto WK, Wong ICK, Leung WK. Statins reduce the progression of non-advanced adenomas to colorectal cancer: A postcolonscopy study in 187 897 patients. *Gut* 2019; pii: gutjnl-2018-317714 [PMID: 30808646 DOI: 10.1136/gutjnl-2018-317714]

70 Dudley JT, Sirotka M, Shenoy M, Pai RK, Roedder S, Chiang AP, Morgan AA, Sarwal MM, Pastrich PJ, Butte AJ. Computational repositioning of the anticonvulsant topiramate for inflammatory bowel disease. *Sci Transl Med* 2011; 3: 96ra76 [PMID: 21849664 DOI: 10.1126/scitranslmed.3002648]

71 Crockett SD, Schectman R, Stürmer T, Kappelman MD. Topiramate use does not reduce flares of inflammatory bowel disease. *Dig Dis Sci* 2014; 59: 1535-1543 [PMID: 24505492 DOI: 10.1007/s10620-014-4040-7]

72 Örqvist AK, Lundholm C, Halfvarson J, Ludvigsson JF, Almqvist C. Fetal and early life antibiotics exposure and very early onset inflammatory bowel disease. *A population-based study. Gut* 2019; 68: 218-225 [PMID: 29231166 DOI: 10.1136/gutjnl-2017-314352]

73 Jännè PA, Mayer RJ. Chemoprevention of colorectal cancer. *N Engl J Med* 2000; 342: 1960-1968 [PMID: 10874065 DOI: 10.1056/nejm200009253424006]

74 Ma Y, Yang W, Song M, Smith-Warner SA, Yang J, Li Y, Ma W, Hu Y, Ongino S, Hu FB, Wen D, Chan AT, Giovannucci EL, Zhang X. Type 2 diabetes and risk of colorectal cancer in two large U.S. prospective cohorts. *Br J Cancer* 2018; 119: 1436-1442 [PMID: 30401889 DOI: 10.1038/s41416-018-0314-4]

75 Liu PH, Wu K, Ng K, Zauber AG, Nguyen LH, Song M, He X, Fuchs CS, Ongino S, Willett WC, Chan AT, Giovannucci EL, Cao Y. Association of Obesity With Risk of Early-Onset Colorectal Cancer Among Women. *JAMA Oncol* 2019; 5: 37-44 [PMID: 30362010 DOI: 10.1001/jamaoncol.2018.4290]

76 He X, Wu K, Ongino S, Giovannucci EL, Chan AT, Song M. Association Between Risk Factors for Colorectal Cancer and Risk of Serrated Polyps and Conventional Adenomas. *Gastroenterology* 2018; 155: 355-373.e18 [PMID: 29702117 DOI: 10.1053/j.gastro.2018.04.019]

77 Hu Y, Ding M, Yuan C, Wu K, Smith-Warner SA, Hu FB, Chan AT, Meyerhardt JA, Ongino S, Fuchs CS, Giovannucci EL, Song M. Association Between Coffee Intake After Diagnosis of Colorectal Cancer and Reduced Mortality. *Gastroenterology* 2018; 154: 916-926.e9 [PMID: 29158191 DOI: 10.1053/j.gastro.2017.11.010]

78 Yang W, Ma Y, Smith-Warner S, Song M, Wu K, Wang M, Chan AT, Ongino S, Fuchs CS, Poylin V, Ng K, Meyerhardt JA, Giovannucci EL, Zhang X. Calcium Intake and Survival after Colorectal Cancer Diagnosis. *Clin Cancer Res* 2019; 25: 1980-1988 [PMID: 30545821 DOI: 10.1158/1078-0432.CCR-18-2965]

79 Song M, Wu K, Meyerhardt JA, Ongino S, Wang M, Fuchs CS, Giovannucci EL, Chan AT. Fiber Intake and Survival After Colorectal Cancer Diagnosis. *JAMA Oncol* 2018; 4: 71-79 [PMID: 29098294 DOI: 10.1001/jamaoncol.2017.3084]

80 de Jong AE, Hendriks YM, Kleibeuker JH, de Boer SY, Cats A, Griffioen G, Nagenast FM, Nelis FG, Rookus MA, Vanen HF. Decrease in mortality in Lynch syndrome families because of surveillance. *Gastroenterology* 2006; 130: 665-671 [PMID: 16530507 DOI: 10.1053/j.gastro.2005.11.032]

81 Engel C, Vanen HF, Sepplöt A, Aretz S, Bigvarmann-Bargeman M, de Boer SY, Buecksch K, Böttler R, Holinski-Feder E, Holzapfel S, Hieneberg R, Jacobs MAJM, Järvinen H, Kloor M, van Kuinzel Doebertz M, Koornstra JJ, van Kruunen M, Langers AM, van de Meeger PC, Morak M, Möülein G, Nagenast FM, Pylvänäinen K, Rahnner N, Renken-Sinisalo L, Sanduleau S, Schaeckert HK, Schmiegel W, Schulmann K, Steinke-Lange V, Strassburg CP, Veitch J, Verhulst ML, de Vos Tot Nederveen Cappel W, Zachariae S, Mecklin JP, Loefler M, German HNPPC Consortium, the Finnish Lynch Syndrome Registry. No Difference in Colorectal Cancer Incidence or Stage at Detection by Colonoscopy Among 3 Different Lynch Syndrome Surveillance Policies. *Gastroenterology* 2018; 155: 1400-1409.e2 [PMID: 30063918 DOI: 10.1053/j.gastro.2018.07.030]

82 Ghorbanoghli Z, Bastianansen BA, Langers AM, Nagenast FM, Foley JW, Hardwick JC, Koornstra JJ, Sanduleau S, de Vos Tot Nederveen Cappel WH, Witteman BJ, Moreau H, Dulker E, Vanen HF. Extracolonic cancer risk in Dutch patients with APC (adenomatous polyposis coli)-associated polyposis. *J
Cheung KS et al. Big Data in gastrointestinal research

Med Genet 2018; 55: 11-14 [PMID: 28490611 DOI: 10.1136/jmedgenet-2017-104545]

Seto WK, Lo YR, Pawlotsky JM, Yuen MF. Chronic hepatitis B virus infection. Lancet 2018; 392: 2313-2324 [PMID: 30496122 DOI: 10.1016/S0140-6736(18)31865-8]

Cheung KS, Seto WK, Lai CL, Yuen MF. Prevention and management of hepatitis B virus reactivation in cancer patients. Hepatol Int 2016; 10: 407-414 [PMID: 26739135 DOI: 10.1007/s12072-015-9692-3]

Cheung KS, Seto WK, Wong DK, Mak LY, Lai CL, Yuen MF. Wisteria floribunda agglutinin-positive human Mac-2 binding protein predicts liver cancer development in chronic hepatitis B patients under antiviral treatment. Oncotarget 2017; 8: 47507-47517 [PMID: 28337906 DOI: 10.18632/oncotarget.17670]

Cheung KS, Seto WK, Wong DK, Lai CL, Yuen MF. Relationship between HBsAg, HBeAg and hepatocellular carcinoma in patients with undetectable HBV DNA under nucleos(t)ide therapy. J Viral Hepat 2017; 24: 654-661 [PMID: 28185363 DOI: 10.1111/jvh.12688]

Wu CY, Lin JT, Ho HJ, Su CW, Lee TY, Wang SY, Wu C, Wu JC. Association of nucleos(t)ide analogue therapy with reduced risk of hepatocellular carcinoma in patients with chronic hepatitis B: A nationwide cohort study. Gastroenterology 2014; 147: 143-151.e1 [PMID: 24704525 DOI: 10.1053/j.gastro.2014.03.048]

Seto WK, Lau EH, Wu JT, Hung IF, Leung WK, Cheung KS, Fung J, Lai CL, Yuen MF. Effects of nucleoside analogue prescription for hepatitis B on the incidence of liver cancer in Hong Kong: A territory-wide ecological study. Aliment Pharmacol Ther 2017; 45: 501-509 [PMID: 27976416 DOI: 10.1111/apt.13895]

Wu CY, Chen YJ, Ho HJ, Hsu YC, Kuo KN, Wu MS, Lin JT. Association between nucleoside analogues and risk of hepatitis B virus-related hepatocellular carcinoma recurrence following liver resection. JAMA 2012; 308: 1906-1914 [PMID: 23162601]

Choi J, Kim HJ, Lee J, Cho S, Ko MJ, Lim YS. Risk of Hepatocellular Carcinoma in Patients Treated With Entecavir vs Tenofovir for Chronic Hepatitis B: A Korean Nationwide Cohort Study. JAMA Oncol 2019; 5: 30-36 [PMID: 30267080 DOI: 10.1001/jamaoncol.2018.4070]

Chen HP, Shih JJ, Chang CC, Chen TT, Lin JT, Wu MS, Lin JH, Wu CY. Metformin decreases hepatocellular carcinoma risk in a dose-dependent manner. Population-based and in vitro studies. Gut 2013; 62: 606-615 [PMID: 22725348 DOI: 10.1136/gutjnl-2011-301708]

Hawkins SS, Gillman MW, Rifas-Shiman SL, Kleinman KP, Mariotti M, Taveras EM. The Linked CENTURY Study: Linking three decades of clinical and public health data to examine disparities in childhood obesity. BMC Pediatr 2016; 16: 32 [PMID: 26961130 DOI: 10.1186/s12876-016-0567-0]

Gevers T, Kugathasan S, D'Costa VC, Van Treuren W, Knight R, Picard A, Yassour M, Dalzell E, Tamura Y, Koren O, Pachter L, Ertel AW, Furey TS, Someren EWJ, Mardis ER, Sabo P, Smolenki J, Abi-Rached L, Yassour M, McElwee J, Cusanovich MA, Conner BA, Anvik T, Skokos A, Verma S, Gouw AS, Fricker MD, Arumugam M, Marra MA, Harper ME, Haas SS, Alm E, Schwalbe D, Ermolaeva M, Usmanova N, Vizcaino RA, McGee Z, Rozenblatt-Rosen O, Reiner A, Geiger T, Hultman E, Karlin S, Spouge JL, Mardis ER, Durbin R, Pachter L, Knight R, Eertman N, Schmid RS, Ertel AW. The treatment-naive microbiome in new-onset Crohn's disease. Cell Host Microbe 2016; 15: 382-392 [PMID: 24629344 DOI: 10.1016/j.chom.2014.02.005]

Harron K, Gilbert R, Cromwell D, van der Meulen J. Linking Data for Mothers and Babies in a De-Identified Electronic Health Data. PLoS One 2016; 11: e0164667 [PMID: 27764135 DOI: 10.1371/journal.pone.0164667]

Wu CY, Wu MS, Kuo KN, Wang CB, Chen YJ, Lin JT. Long-term peptic ulcer rebleeding risk estimation in patients undergoing haemodialysis: A 10-year nationwide cohort study. Gut 2011; 60: 1038-1042 [PMID: 21266725 DOI: 10.1136/gut.2010.224329]

Kim MS, Song HJ, Lee J, Yang BR, Choi NK, Park BJ. Effectiveness and Safety of Clopidogrel Co-administered With Statins and Proton Pump Inhibitors: A Korean National Health Insurance Database Study. Clin Pharmacol Ther 2019 [PMID: 30648733 DOI: 10.1002/cpt.1361]

Song EM, Lee HS, Kim YJ, Oh EH, Ham NS, Kim J, Hwang SW, Park SH, Yang DH, Ye BD, Byeon JS, Myung SJ, Yang SK. Incidence and clinical impact of perianal disease in patients with ulcerative colitis: A nationwide population-based study. J Gastroenterol Hepatol 2018 [PMID: 30549125 DOI: 10.1111/jgh.14555]

Chang K, Lee HS, Kim YJ, Kim SO, Kim SH, Lee SH, Song EM, Hwang SW, Park SH, Yang DH, Ye BD, Byeon JS, Myung SJ, Yang SK. Increased risk of Herpes Zoster Infection in Patients With Inflammatory Bowel Diseases in Korea. Clin Gastroenterol Hepatol 2018; 16: 1928-1936.e2 [PMID: 29857190 DOI: 10.1016/j.cgh.2018.05.024]

Myrelid P, Landerholm K, Nordenval C, Pinkney TD, Andersson RE. Appendectomy and the Risk of Colonectomy in Ulcerative Colitis: A National Cohort Study. Am J Gastroenterol 2017; 112: 1311-1319 [PMID: 28653667 DOI: 10.1038/ajg.2017.183]

Hou JK, Chang M, Nguyen T, Kramer JR, Richardson P, Samsgry S, D’Avolio LW, El-Serag HB. Automated identification of surveillance colonscopy in inflammatory bowel disease using natural language processing. Dig Dis Sci 2013; 58: 936-941 [PMID: 23086115 DOI: 10.1007/s10620-012-2433-8]

Cheung KS, Chen L, Seto WK, Leung WK. Epidemiology, characteristics and survival of post-colonoscopy colorectal cancer in Asia: A population-based study. J Gastroenterol Hepatol 2019 [PMID: 30932240 DOI: 10.1111/jgh.14674]

Ji J, Sundaquist K, Sundquist J. A population-based study of hepatitis D virus as potential risk factor for hepatocellular carcinoma. J Natl Cancer Inst 2012; 104: 790-792 [PMID: 22423008 DOI: 10.1093/jnci/djs168]
