WJG

World Journal of
Gastroenterology

Submit a Manuscript: https://www.f6publishing.com

World J Gastroenterol 2019 June 28; 25(24): 2990-3008

DOI: 10.3748/wjg.v25.i24.2990

ISSN 1007-9327 (print) ISSN 2219-2840 (online)

REVIEW

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.

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

Ka-Shing Cheung, Wai K Leung, Wai-Kay Seto, Department of Medicine, The University of Hong Kong, Queen Mary Hospital, Hong Kong, China

Ka-Shing Cheung, Wai-Kay Seto, Department of Medicine, The University of Hong Kong-Shenzhen Hospital, Shenzhen 518053, Guangdong Province, China

Corresponding author: Wai-Kay Seto, FRCP (C), MBBS, MD, MRCP, Associate Professor, Department of Medicine, The University of Hong Kong, Queen Mary Hospital, 102 Pokfulam Road, Hong Kong, China. wkseto@hku.hk
Telephone: +86-852-22553994
Fax: +86-852-28725828

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

Article in press: April 29, 2019
Published online: June 28, 2019

P-Reviewer: Adibi P, Amornyotin S, Contini S
S-Editor: Yan JP
L-Editor: Filipodia
E-Editor: Zhang YL

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[2] 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[2].

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[9]. 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[10]. 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 prognostic model be developed, which ultimately guides a targeted approach for selecting treatment regimens for individual patients (*i.e.*, precision or personalized medicine)[9,11,12].

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[13]. 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[14]). 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[15]. 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”[16]. 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)][17,18], 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[19]. 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[20]). The sources are not limited to medical literature and clinical notes, but also product labelling, social media and web search logs[21,22].

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[23]. 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 effects due to time 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 = 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[24]. The possibility of causality can also be strengthened *via* the fulfilment of the Bradford Hill criteria[25]. Second, data validity concerning the accuracy of diagnosis codes (*e.g.*, International Classification of Diseases) in electronic databases has been challenged[26]. 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[27]. 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[28]. 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[26] (*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][29]. As an example, free-text searches outperformed discharge diagnosis coding in the detection of postoperative complications[30]. 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[31]. 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 multi-centre, 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[32]. 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	
Clinical data readily available with minimal resources required	
Can study rare exposures	
Can study rare events	
Can study long-term effects	
Real-world data	
Large sample size	
Subgroup analysis	
Sensitivity analysis	
Interaction of different variables	
Adjustment of outcome to a multitude of risk factors	
Precise estimation of effect size	
Reliable capture of small variations in incidence or disease flare	
No selection bias if *n* = 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
	Text mining or natural language processing of unstructured data
Incomplete capture of variables or unavailability of certain diagnosis codes	Surrogate markers (*e.g.*, COPD for smoking, alcohol-related diseases for alcoholism)
	Inclusion of a large set of measured variables
	Text mining or natural language processing of unstructured data
Privacy	De-identification of individuals
	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	Inclusion of a large set of measured variables
	Inclusion of RCT datasets with extensive collection of data and outcomes for trial participants or linkage with other data sources
	Fulfilment of Bradford Hill criteria
Reverse causality/protopathic bias (outcome of interest leads to exposure of interest)	Cohort study design instead of case-control study design
Example: Early symptoms of undiagnosed GC leads to PPI use, rather than PPIs cause GC	Excluding prescriptions of drugs of interest (*e.g.*, PPIs) within a certain period (*e.g.*, 6 mo) before development of the outcome of interest (*e.g.*, gastric cancer)
Selection bias	Encompassing entire study population (*n* = all)
Indication bias (or confounding by indication/disease severity)	Balance of patient characteristics, in particular comorbidities that are indications for a certain treatment (*e.g.*, PS matching of a large set of measured variables)
	Negative control exposure
Confounding by functional status and cognitive impairment	Balance of patient characteristics, in particular comorbidities that can affect functional and cognitive status (*e.g.*, PS matching)
Healthy user bias / adherer bias (individuals who are more health conscious tend to have better health outcomes)	Adjustment for other lifestyle factors – text mining or natural language processing of unstructured data
Immortal time bias (arises when the study outcome cannot occur during a period of follow-up due to study design)	Landmark analysis
	Analysis using time varying covariates
Ascertainment bias / surveillance bias / detection bias (differential degree of surveillance or screening for the outcome among exposed and unexposed individuals) Example: PPI users may undergo upper endoscopy more frequently than non-PPI users, and hence more GC detected in PPI users	Selection of an unexposed group with a similar likelihood of screening/testing
	Selection of an outcome that are likely to be diagnosed equally in exposed and control groups
	Adjustment for the surveillance rate
Access to healthcare	Stratified analysis according to patients' residential regions (*e.g.*, rural *vs* urban), socioeconomic status, immigration status, race/ethnicity, institutional factors (*e.g.*, restrictive formularies)
Selective prescription and treatment in frail and very sick patients	PS methodology (trimming of areas of non-overlap, PS matching, PS by treatment interaction)

COPD: Chronic pulmonary obstructive disease; RCT: Randomized controlled trial; GC: Gastric cancer; PPI: Proton pump inhibitor; PS: Propensity score.

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"[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[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[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[35]. 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*[38] on the effect of tissue plasminogen activator on in-hospital mortality. Table 2 summarizes the major advantages of PS methodologies.

EXAMPLES OF GASTROINTESTINAL DISEASE RESEARCHE 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[39]. Around two-thirds of patients have gastric cancer diagnosed at an advanced stage, rendering curative surgery impossible[40,41]. Infection by *Helicobacter pylori (H. pylori)*, a class I human carcinogen[42], confers a two- to three-fold increase in gastric cancer risk[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[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[46]. Based on a territory-wide public healthcare database in Hong Kong

Table 2 Advantages of propensity score methodology

Advantages	Remarks
Addressing "curse of dimensionality" when EPV < 10	Traditional multivariable regression models yield similar results if EPV ≥ 10
Recognition of subjects with absolute indications (or contraindications) of an intervention	Exclusion of areas of non-overlap of the PS distribution between exposed and unexposed groups to ensure comparability
Identification of PS interaction with treatment	Variation of effectiveness of an intervention according to indications (PS) may only be identified *via* stratified analysis by PS

EPV: Events per variable; PS: Propensity score.

called the Clinical Data Analysis and Reporting System, *H. pylori* eradication therapy was beneficial even in older age groups (≥ 60 years)[47]. 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[48]. Long-term aspirin use further reduced gastric cancer risk in patients who had received *H. pylori* eradication therapy[49]. 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[50].

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[51], which is otherwise difficult to be addressed by RCTs[52]. This finding was echoed by another nationwide study[53]. 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[54]. 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[55]. 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[56]. In addition, racial/ethnic minorities had a 40%-50% increase in gastric cancer risk compared with the Hispanic and white populations[57].

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[5]. 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%)[58]. 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[59]. This harmful effect was particularly prominent for PPIs with high CYP2C19 inhibitory potential[60]. 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[61]. 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[62]. 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 involving more than 3 million NOAC users[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

Gastric cancer

Country/Region	Database	Area of research	Sample size	Design, statistical methods and 3V	Application
Taiwan, China	Taiwan National Health Insurance Database (NHID)	GC Wu *et al*[46], 2009	80255	Nationwide retrospective cohort study Comparison with general population to derive SIR Volume, Velocity and Variety	Early *vs* late *H. pylori* eradication on GC risk
		GC Wu *et al*[48], 2010	52161	Nationwide retrospective cohort study Comparison with general population to derive SIR Volume, Velocity and Variety	Association between NSAIDs and GC
Hong Kong, China	Clinical Data Analysis and Reporting System (CDARS)	GC Cheung *et al*[51], 2018	63397	Territory-wide retrospective cohort study PS regression adjustment Volume, Velocity and Variety	Association between PPIs and GC
		GC Cheung *et al*[49], 2018	63605	Territory-wide retrospective cohort study PS regression adjustment Volume, Velocity and Variety	Association between aspirin and GC
		GC Leung *et al*[47], 2018	63397	Territory-wide retrospective cohort study Comparison with general population to derive SIR Volume, Velocity and Variety	Effect of *H. pylori* eradication among different age groups
		GC Cheung *et al*[50], 2018	7266	Territory-wide retrospective cohort study PS regression adjustment Sensitivity analysis: PS weighting by IPTW and PS matching Volume, Velocity and Variety	Association between metformin and GC
Sweden	Swedish Cancer Registry Swedish Prescribed Drug Registry	GC Brusselaers *et al*[53], 2017	797067	Nationwide retrospective cohort study Comparison with general population to derive SIR Volume, Velocity and Variety	Association between PPIs and GC
		GC Doorakkers *et al*[45], 2018	95176	Nationwide retrospective cohort study Comparison with general population to derive SIR	Effect of *H. pylori* eradication on GC risk

				Volume, Velocity and Variety	
United States	Kaiser Permanente (KP)	GC Schneider *et al*[55], 2016	61684	Retrospective cohort study Volume, Velocity and Variety	Association between different PPIs and GC

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; GC: Gastric cancer; SIR: Standardized incidence ratio; *H. pylori*: *Helicobacter pylori*; NSAIDs: Non-steroidal anti-inflammatory drugs; PS: Propensity score; PPIs: Proton pump inhibitors; IPTW: Inverse probability of treatment weighting.

disease with numerous predictive factors. Machine learning algorithms are particularly useful in deriving predictive models, including risk factors[64], disease outcomes[65] and treatment responses[66,67], 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 $\alpha_4\beta_7$ 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 antisense oligonucleotide showing a promising effect in a phase II trial in Crohn's disease[68], was prematurely terminated in the phase III program[69]. The results for secukinumab, 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[14], 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*[70]. 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[70]. A technique, called "signature inversion"[16], 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[70]. The potential role of topiramate, however, was later refuted by a retrospective cohort study[71], 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[20]. As the electronic coding of arthralgia is not commonly performed in gastroenterology practices, Cai *et al*[20] 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[72]. 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[39]. As a period of 10 years is required for the development of the adenoma-carcinoma sequence[73], 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, NHSII and HPFS cohorts. Type II diabetes mellitus was associated with a 1.4-fold increase in CRC risk[74]. A positive association between obesity and early-onset CRC also existed among women[75]. 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[76]. 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

Gastrointestinal bleeding and/or proton pump inhibitors					
Country/Region	**Database**	**Area of research**	**Sample size**	**Design, statistical methods and 3V**	**Application**
Taiwan, China	Taiwan National Health Insurance Database (NHID)	PUD Wu *et al*[58], 2009	403567	Nationwide retrospective cohort study Volume, Velocity and Variety	Effect of *H. pylori* therapy and PPIs on PUD
		PUD Wu *et al*[95], 2011	32235	Nationwide retrospective cohort study Volume, Velocity and Variety	Risk of rebleeding from PUD in ESRD patients
		PPIs Wu *et al*[59], 2010	6552	Nationwide retrospective cohort study Volume, Velocity and Variety	Effect of clopidogrel and PPIs on ACS
South Korea	Korean Health Insurance Review and Assessment Service (HIRA)	PPIs Kim *et al*[96], 2019	59233	Nationwide retrospective cohort study Volume, Velocity and Variety	Effect of PPIs on thrombotic risk
Hong Kong, China	Clinical Data Analysis and Reporting System (CDARS)	Dabigatran Chan *et al*[62], 2015	5041	Territory-wide retrospective cohort study Volume, Velocity and Variety	Risk factors for dabigatran-associated gastrointestinal bleeding

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

Inflammatory bowel disease					
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 Song *et al*[97], 2018	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 Chang *et al*[98], 2018	38039	Nationwide retrospective cohort study to compare IBD patients with general population to derive SIR Hospital based nested case-control study Volume, Velocity and Variety	Association between IBD and herpes zoster infection
Sweden	Swedish Patient Registry	UC Myrelid *et al*[99], 2017	63711	Nationwide retrospective cohort study Volume, Velocity and Variety	Association between appendectomy and UC
	Swedish Medical Birth Register (child-mother link) Swedish Multigeneration Register (child-father link) Swedish Prescribed Drug Register National Patient Register	IBD Ortqvist *et al*[72], 2019	827,239 children born between 2006 and 2013	Nationwide prospective population-based register study Volume, Velocity and Variety	Association between maternal exposure to antibiotics during pregnancy and very early onset IBD in adulthood
United States	NCBI Gene Expression Omnibus (GEO)	IBD Dudley *et al*[70], 2011	n.a.	Signature inversion study Volume, Velocity and Variety	Topiramate as a potential therapeutic agent against IBD
United States	n.a.	IBD Cai *et al*[20], 2018	1585	Retrospective cohort study Natural language processing Volume, Velocity and Variety	Association between arthralgia and biologics (anti-TNF *vs* vedolizumab)
n.a	International IBD Genetics Consortium's Immunochip project	IBD Wei *et al*[64], 2013	53279	Machine learning algorithm Volume, Velocity and Variety	Predictors of IBD
United States	n.a.	IBD Hou *et al*[100], 2013	575 colonoscopy reports	Retrospective cohort study Natural language processing Volume, Velocity and Variety	Differentiation of surveillance from non-surveillance colonoscopy
United States	n.a.	IBD Waljee *et al*[66], 2017	1080	Retrospective cohort study Random Forest machine learning algorithm	Prediction of IBD remission in thiopurine users
United States	n.a.	IBD Waljee *et al*[65], 2017	20368	Retrospective cohort study Random Forest machine learning algorithm	Prediction of hospitalization and outpatient steroid use
n.a.	Phase 3 clinical trial data	IBD Waljee *et al*[67], 2018	491	Retrospective cohort study Random Forest machine learning algorithm	Prediction of steroid-free endoscopic remission with vedolizumab in UC

		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; UC: Ulcerative colitis; IBD: Inflammatory bowel disease; SIR: Standardized incidence ratio; anti-TNF: anti-tumour necrosis factor.

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*[17] 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[18].

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[92]. 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[72]. One possible mechanism is *via* the alteration of the gut microbiome[93]. 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[94]. 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

Colorectal cancer

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 Volume, Velocity and Variety	Epidemiology, characteristics, risk factors and prognosis of postcolonoscopy Colorectal cancer in Asians
		CRC Cheung *et al*[69], 2019	187897	Territory-wide retrospective cohort study PS matching Volume, Velocity and Variety	Association between statins and CRC
United States	Nurses' Health Study II (NHSII) Health Professionals Follow-up Study (HPFS)	CRC Ma *et al*[74], 2018	134763	Prospective cohort study Volume and Variety	Association between DM and CRC
	Nurses' Health Study (NHS) Health Professionals Follow-up Study (HPFS)	CRC Yang *et al*[78], 2018 Hu *et al*[77], 2018 Song *et al*[79], 2018	1660 1599 1575	Prospective cohort study Volume and Variety	Effect of calcium intake, coffee and fibre on survival after CRC diagnosis
	Nurses' Health Study (NHS) Nurses' Health Study II (NHSII) Health Professionals Follow-up Study (HPFS)	CRC He *et al*[76], 2018 de Jong *et al*[80], 2006	141143	Prospective cohort study Volume and Variety	Risk factors of serrated polyps and conventional adenomas
	Nurses' Health Study II (NHSII)	CRC Liu *et al*[75], 2018	85256	Prospective cohort study Volume and Variety	Association between obesity and CRC
Netherlands	Dutch Lynch syndrome Registry	Various cancers including CRC	2788	Retrospective cohort study Volume, Velocity and Variety	Decrease in CRC-related mortality in Lynch syndrome families by surveillance
Netherlands, Germany, Finland	Dutch Lynch syndrome Registry German HNPCC Consortium Finland	CRC Engel *et al*[81], 2018	2747 patients with 16327 colonoscopies	Retrospective cohort study 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

Hepatocellular carcinoma

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

	Taiwan National Health Insurance Database (NHID)	HCC Wu *et al*[89], 2012	4569	Nationwide retrospective cohort study Volume, Velocity and Variety	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*[91], 2013	292290	Nationwide case-control study Volume, Velocity and Variety	Association between DM and HCC
	Taiwan National Health Insurance Database (NHID)	HCC Wu *et al*[87], 2014	43190	Nationwide retrospective cohort study PS matching Volume, Velocity and Variety	Association between NA therapy and HCC among CHB patients
China	The Cancer Genome Atlas (TCGA) database Connectivity Map (CMap)	HCC Wang *et al*[18], 2016	n.a.	Signature inversion study Volume, Velocity and Variety	Anti-cancer effect of prenylamine on HCC
South Korea	Korean Health Insurance Review and Assessment Service (HIRA)	HCC Choi *et al*[90], 2018	24156	Nationwide retrospective cohort study Volume, Velocity and Variety	Difference between tenofovir and entecavir on reducing HCC risk
Hong Kong, China	Clinical Data Analysis and Reporting System (CDARS)	HCC Seto *et al*[88], 2017	Entire Hong Kong population between 1999 and 2012	Territory-wide retrospective cohort study Volume, Velocity and Variety	Association between NA therapy and HCC among CHB patients
Sweden	Swedish Cancer Registry Swedish Patient Registry	HCC Ji *et al*[102], 2012	9160 CHB patients	Nationwide retrospective cohort study Comparison with general population to derive SIR Volume, Velocity and Variety	Association between concomitant HBV/HDV infection and 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.

REFERENCES

1 **Lohr S**. The Origins of 'Big Data': An Etymological Detective Story [cited 25 January 2019]. Available from: https://bits.blogs.nytimes.com/2013/02/01/the-origins-of-big-data-an-etymological-detective-story

2 **Kitchin R**, McArdle 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 Comission**. Study on Big Data in Public Health, Telemedicine and Healthcare. December. 2016 [DOI: 10.2875/734795]

7 **Alonso SG**, de la Torre Díez I, Rodrigues JJPC, Hamrioui S, López-Coronado 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 **Olivera P**, Danese S, Jay N, Natoli G, Peyrin-Biroulet L. Big data in IBD: A look into the future. *Nat Rev Gastroenterol Hepatol* 2019 [PMID: 30659247 DOI: 10.1038/s41575-019-0102-5]

10 **Mirkov MU**, 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, Fosler-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*

Assoc 2014; **21**: 221-230 [PMID: 24201027 DOI: 10.1136/amiajnl-2013-001935]

12 **Zuo T**, Kamm MA, Colombel JF, Ng SC. Urbanization and the gut microbiota in health and inflammatory bowel disease. *Nat Rev Gastroenterol Hepatol* 2018; **15**: 440-452 [PMID: 29670252 DOI: 10.1038/s41575-018-0003-z]

13 **Schuhmacher A**, Gassmann O, Hinder M. Changing R and amp;D models in research-based pharmaceutical companies. *J Transl Med* 2016; **14**: 105 [PMID: 27118048 DOI: 10.1186/s12967-016-0838-4]

14 **Hueber W**, Sands BE, Lewitzky S, Vandemeulebroecke M, Reinisch W, Higgins PD, Wehkamp J, Feagan BG, Yao MD, Karczewski M, Karczewski J, Pezous N, Bek S, Bruin G, Mellgard B, Berger C, Londei M, Bertolino AP, Tougas G, Travis SP; Secukinumab in Crohn's Disease Study Group. Secukinumab, a human anti-IL-17A monoclonal antibody, for moderate to severe Crohn's disease: Unexpected results of a randomised, double-blind placebo-controlled trial. *Gut* 2012; **61**: 1693-1700 [PMID: 22595313 DOI: 10.1136/gutjnl-2011-301668]

15 **Denny JC**, Van Driest SL, Wei WQ, Roden DM. The Influence of Big (Clinical) Data and Genomics on Precision Medicine and Drug Development. *Clin Pharmacol Ther* 2018; **103**: 409-418 [PMID: 29171014 DOI: 10.1002/cpt.951]

16 **Li J**, Zheng S, Chen B, Butte AJ, Swamidass SJ, Lu Z. A survey of current trends in computational drug repositioning. *Brief Bioinform* 2016; **17**: 2-12 [PMID: 25832646 DOI: 10.1093/bib/bbv020]

17 **Chen MH**, Yang WL, Lin KT, Liu CH, Liu YW, Huang KW, Chang PM, Lai JM, Hsu CN, Chao KM, Kao CY, Huang CY. Gene expression-based chemical genomics identifies potential therapeutic drugs in hepatocellular carcinoma. *PLoS One* 2011; **6**: e27186 [PMID: 22087264 DOI: 10.1371/journal.pone.0027186]

18 **Wang J**, Li M, Wang Y, Liu X. Integrating subpathway analysis to identify candidate agents for hepatocellular carcinoma. *Onco Targets Ther* 2016; **9**: 1221-1230 [PMID: 27022281 DOI: 10.2147/OTT.S97211]

19 **Graul AI**, Cruces E, Stringer M. The year's new drugs &amp; biologics, 2013: Part I. *Drugs Today (Barc)* 2014; **50**: 51-100 [PMID: 24524105 DOI: 10.1358/dot.2014.50.1.2116673]

20 **Cai T**, Lin TC, Bond A, Huang J, Kane-Wanger G, Cagan A, Murphy SN, Ananthakrishnan AN, Liao KP. The Association Between Arthralgia and Vedolizumab Using Natural Language Processing. *Inflamm Bowel Dis* 2018; **24**: 2242-2246 [PMID: 29846617 DOI: 10.1093/ibd/izy127]

21 **Harpaz R**, Callahan A, Tamang S, Low Y, Odgers D, Finlayson S, Jung K, LePendu P, Shah NH. Text mining for adverse drug events: The promise, challenges, and state of the art. *Drug Saf* 2014; **37**: 777-790 [PMID: 25151493 DOI: 10.1007/s40264-014-0218-z]

22 **Wang G**, Jung K, Winnenburg R, Shah NH. A method for systematic discovery of adverse drug events from clinical notes. *J Am Med Inform Assoc* 2015; **22**: 1196-1204 [PMID: 26232442 DOI: 10.1093/jamia/ocv102]

23 **Genta RM**, Sonnenberg A. Big data in gastroenterology research. *Nat Rev Gastroenterol Hepatol* 2014; **11**: 386-390 [PMID: 24594912 DOI: 10.1038/nrgastro.2014.18]

24 **Hsing AW**, Ioannidis JP. Nationwide Population Science: Lessons From the Taiwan National Health Insurance Research Database. *JAMA Intern Med* 2015; **175**: 1527-1529 [PMID: 26192815 DOI: 10.1001/jamainternmed.2015.3540]

25 **Cheung KS**, Leung WK. Response to letter to the editor by Moayyedi *et al*. *Gut* 2018; **pii**: gutjnl-2018-317127 [PMID: 30121628 DOI: 10.1136/gutjnl-2018-317127]

26 **Schneeweiss S**, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. *J Clin Epidemiol* 2005; **58**: 323-337 [PMID: 15862718 DOI: 10.1016/j.jclinepi.2004.10.012]

27 **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 [PMID: 20473199 DOI: 10.1097/MLR.0b013e3181dbebe3]

28 **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]

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

30 **Murff HJ**, FitzHenry F, Matheny ME, Gentry N, Kotter KL, Crimin K, Dittus RS, Rosen AK, Elkin 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]

31 **Vayena E**, Salathé M, Madoff LC, Brownstein JS. Ethical challenges of big data in public health. *PLoS Comput Biol* 2015; **11**: e1003904 [PMID: 25664461 DOI: 10.1371/journal.pcbi.1003904]

32 **D'Agostino RB**. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat Med* 1998; **17**: 2265-2281 [PMID: 9802183]

33 **Schulte PJ**, Mascha EJ. Propensity Score Methods: Theory and Practice for Anesthesia Research. *Anesth Analg* 2018; **127**: 1074-1084 [PMID: 29750691 DOI: 10.1213/ANE.0000000000002920]

34 **Peduzzi P**, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol* 1996; **49**: 1373-1379 [PMID: 8970487 DOI: 10.1016/S0895-4356(96)00236-3]

35 **Glynn RJ**, Schneeweiss S, Stürmer T. Indications for propensity scores and review of their use in pharmacoepidemiology. *Basic Clin Pharmacol Toxicol* 2006; **98**: 253-259 [PMID: 16611199 DOI: 10.1111/j.1742-7843.2006.pto_293.x]

36 **Shah BR**, Laupacis A, Hux JE, Austin PC. Propensity score methods gave similar results to traditional regression modeling in observational studies: A systematic review. *J Clin Epidemiol* 2005; **58**: 550-559 [PMID: 15878468 DOI: 10.1016/j.jclinepi.2004.10.016]

37 **Stürmer T**, Joshi M, Glynn RJ, Avorn J, Rothman KJ, Schneeweiss S. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *J Clin Epidemiol* 2006; **59**: 437-447 [PMID: 16632131 DOI: 10.1016/j.jclinepi.2005.07.004]

38 **Kurth T**, Walker AM, Glynn RJ, Chan KA, Gaziano JM, Berger K, Robins JM. Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *Am J Epidemiol* 2006; **163**: 262-270 [PMID: 16371515 DOI: 10.1093/aje/kwj047]

39 **Global Burden of Disease Cancer Collaboration**. Fitzmaurice C, Allen C, Barber RM, Barregard L,

Bhutta ZA, Brenner H, Dicker DJ, Chimed-Orchir 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, Kutz M, Lan Q, Larson HJ, Liang X, Lim SS, Lopez AD, MacIntyre MF, Marczak L, Marquez N, Mokdad 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, Zoeckler L, Abd-Allah F, Ahmed MB, Alabed S, Alam NK, Aldhahri SF, Alem G, Alemayohu MA, Ali R, Al-Raddadi R, Amare A, Amoako Y, Artaman A, Asayesh H, Atnafu N, Awasthi A, Saleem HB, Barac A, Bedi N, Bensenor I, Berhane A, Bernabé E, Betsu B, Binagwaho A, Boneya D, Campos-Nonato I, Castañeda-Orjuela C, Catalá-López F, Chiang P, Chibueze C, Chitheer 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, Gopalani S, Hailu A, Horino M, Horita N, Husseini A, Huybrechts I, Inoue M, Islami F, Jakovljevic M, James S, Javanbakht M, Jee SH, Kasaeian A, Kedir MS, Khader YS, Khang YH, Kim D, Leigh J, Linn S, Lunevicius R, El Razek HMA, Malekzadeh R, Malta DC, Marcenes W, Markos D, Melaku YA, Meles KG, Mendoza W, Mengiste DT, Meretoja TJ, Miller TR, Mohammad KA, Mohammadi A, Mohammed S, Moradi-Lakeh M, Nagel G, Nand D, Le Nguyen Q, Nolte S, Ogbo FA, Oladimeji KE, Oren E, Pa M, Park EK, Pereira DM, Plass D, Qorbani M, Radfar A, Rafay A, Rahman M, Rana SM, Søreide K, Satpathy M, Sawhney M, Sepanlou SG, Shaikh MA, She J, Shiue I, Shore HR, Shrime MG, So S, Soneji S, Stathopoulou V, Stroumpoulis K, Sufiyan MB, Sykes BL, Tabarés-Seisdedos R, Tadese F, Tedla BA, Tessema GA, Thakur JS, Tran BX, Ukwaja KN, Uzochukwu BSC, Vlassov VV, Weiderpass E, Wubshet Terefe M, Yebyo HG, Yimam HH, Yonemoto N, Younis MZ, Yu C, Zaidi Z, Zaki 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 [PMID: 27918777 DOI: 10.1001/jamaoncol.2016.5688]

40 **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 [PMID: 23102520 DOI: 10.1016/j.ctrv.2012.09.007]

41 **Van Cutsem E**, Sagaert X, Topal B, Haustermans K, Prenen H. Gastric cancer. *Lancet* 2016; **388**: 2654-2664 [PMID: 27156933 DOI: 10.1016/S0140-6736(16)30354-3]

42 **Infection with Helicobacter pylori**. *IARC Monogr Eval Carcinog Risks Hum* 1994; **61**: 177-240 [PMID: 7715070]

43 **Cavaleiro-Pinto M**, Peleteiro B, Lunet N, Barros H. Helicobacter pylori infection and gastric cardia cancer: Systematic review and meta-analysis. *Cancer Causes Control* 2011; **22**: 375-387 [PMID: 21184266 DOI: 10.1007/s10552-010-9707-2]

44 **Cheung KS**, Leung WK. Risk of gastric cancer development after eradication of Helicobacter pylori. *World J Gastrointest Oncol* 2018; **10**: 115-123 [PMID: 29770171 DOI: 10.4251/wjgo.v10.i5.115]

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

46 **Wu CY**, Kuo KN, Wu MS, Chen YJ, Wang CB, Lin JT. Early Helicobacter pylori eradication decreases risk of gastric cancer in patients with peptic ulcer disease. *Gastroenterology* 2009; **137**: 1641-8.e1-2 [PMID: 19664631 DOI: 10.1053/j.gastro.2009.07.060]

47 **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 [PMID: 29550592 DOI: 10.1053/j.gastro.2018.03.028]

48 **Wu CY**, Wu MS, Kuo KN, Wang CB, Chen YJ, Lin JT. Effective reduction of gastric cancer risk with regular use of nonsteroidal anti-inflammatory drugs in Helicobacter pylori-infected patients. *J Clin Oncol* 2010; **28**: 2952-2957 [PMID: 20479409 DOI: 10.1200/JCO.2009.26.0695]

49 **Cheung KS**, Chan EW, Wong AYS, Chen L, Seto WK, Wong ICK, Leung WK. Aspirin and Risk of Gastric Cancer After Helicobacter pylori Eradication: A Territory-Wide Study. *J Natl Cancer Inst* 2018; **110**: 743-749 [PMID: 29361002 DOI: 10.1093/jnci/djx267]

50 **Cheung KS**, Chan EW, Wong AYS, Chen L, Seto WK, Wong ICK, Leung WK. Metformin Use and Gastric Cancer Risk in Diabetic Patients After Helicobacter pylori Eradication. *J Natl Cancer Inst* 2019; **111**: 484-489 [PMID: 30329127 DOI: 10.1093/jnci/djy144]

51 **Cheung KS**, Chan EW, Wong AYS, Chen L, Wong ICK, Leung WK. Long-term proton pump inhibitors and risk of gastric cancer development after treatment for Helicobacter pylori: A population-based study. *Gut* 2018; **67**: 28-35 [PMID: 29089382 DOI: 10.1136/gutjnl-2017-314605]

52 **Cheung KS**, Leung WK. Long-term use of proton-pump inhibitors and risk of gastric cancer: A review of the current evidence. *Therap Adv Gastroenterol* 2019; **12**: 1756284819834511 [PMID: 30886648 DOI: 10.1177/1756284819834511]

53 **Brusselaers N**, Wahlin K, Engstrand L, Lagergren J. Maintenance therapy with proton pump inhibitors and risk of gastric cancer: A nationwide population-based cohort study in Sweden. *BMJ Open* 2017; **7**: e017739 [PMID: 29084798 DOI: 10.1136/bmjopen-2017-017739]

54 **Cheung KS**, Leung WK. Modification of gastric cancer risk associated with proton pump inhibitors by aspirin after Helicobacter pylori eradication. *Oncotarget* 2018; **9**: 36891-36893 [PMID: 30651922 DOI: 10.18632/oncotarget.26382]

55 **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 [PMID: 26541643 DOI: 10.1111/apt.13450]

56 **Reddy KM**, Chang JI, 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 [PMID: 27317852 DOI: 10.1016/j.cgh.2016.05.045]

57 **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 [PMID: 27939654 DOI: 10.1016/j.cgh.2016.11.033]

58 **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 [PMID: 19264578 DOI: 10.1016/j.cgh.2008.12.029]

59 **Wu CY**, Chan FK, Wu MS, Kuo KN, Wang CB, Tsao CR, Lin JT. Histamine2-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 IC. 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, Daugherty 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, Mentch 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: 23731541 DOI: 10.1016/j.ajhg.2013.05.002]

65 **Waljee AK**, Lipson R, Wiitala WL, Zhang Y, Liu B, Zhu J, Wallace B, Govani SM, Stidham 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/izx007]

66 **Waljee AK**, Sauder K, Patel A, Segar S, Liu B, Zhang Y, Zhu J, Stidham RW, Balis U, Higgins PDR. Machine Learning Algorithms for Objective Remission and Clinical Outcomes with Thiopurines. *J Crohns Colitis* 2017; **11**: 801-810 [PMID: 28333183 DOI: 10.1093/ecco-jcc/jjx014]

67 **Waljee AK**, Liu B, Sauder K, Zhu J, Govani SM, Stidham 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 postcolonoscopy study in 187 897 patients. *Gut* 2019; **pii**: gutjnl-2018-317714 [PMID: 30808646 DOI: 10.1136/gutjnl-2018-317714]

70 **Dudley JT**, Sirota M, Shenoy M, Pai RK, Roedder S, Chiang AP, Morgan AA, Sarwal MM, Pasricha 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: 24504592 DOI: 10.1007/s10620-014-3040-7]

72 **Örtqvist 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: 29321166 DOI: 10.1136/gutjnl-2017-314352]

73 **Jänne PA**, Mayer RJ. Chemoprevention of colorectal cancer. *N Engl J Med* 2000; **342**: 1960-1968 [PMID: 10874065 DOI: 10.1056/nejm200006293422606]

74 **Ma Y**, Yang W, Song M, Smith-Warner SA, Yang J, Li Y, Ma W, Hu Y, Ogino 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, Ogino 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: 30326010 DOI: 10.1001/jamaoncol.2018.4280]

76 **He X**, Wu K, Ogino 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, Ogino 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, Ogino 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, Ogino 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.2017.3684]

80 **de Jong AE**, Hendriks YM, Kleibeuker JH, de Boer SY, Cats A, Griffioen G, Nagengast FM, Nelis FG, Rookus MA, Vasen 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**, Vasen HF, Seppälä T, Aretz S, Bigirwamungu-Bargeman M, de Boer SY, Bucksch K, Büttner R, Holinski-Feder E, Holzapfel S, Hüneburg R, Jacobs MAJM, Järvinen H, Kloor M, von Knebel Doeberitz M, Koornstra JJ, van Kouwen M, Langers AM, van de Meeberg PC, Morak M, Möslein G, Nagengast FM, Pylvänäinen K, Rahner N, Renkonen-Sinisalo L, Sanduleanu S, Schackert HK, Schmiegel W, Schulmann K, Steinke-Lange V, Strassburg CP, Vecht J, Verhulst ML, de Vos Tot Nederveen Cappel W, Zachariae S, Mecklin JP, Loeffler M; German HNPCC Consortium, the Dutch Lynch Syndrome Collaborative Group, and the Finnish Lynch Syndrome Registry. No Difference in Colorectal Cancer Incidence or Stage at Detection by Colonoscopy Among 3 Countries With Different Lynch Syndrome Surveillance Policies. *Gastroenterology* 2018; **155**: 1400-1409.e2 [PMID: 30063918 DOI: 10.1053/j.gastro.2018.07.030]

82 **Ghorbanoghli Z**, Bastiaansen BA, Langers AM, Nagengast FM, Poley JW, Hardwick JC, Koornstra JJ, Sanduleanu S, de Vos Tot Nederveen Cappel WH, Witteman BJ, Morreau H, Dekker E, Vasen HF. Extracolonic cancer risk in Dutch patients with APC (adenomatous polyposis coli)-associated polyposis. *J*

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

83 **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]

84 **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]

85 **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: 28537900 DOI: 10.18632/oncotarget.17670]

86 **Cheung KS**, Seto WK, Wong DK, Lai CL, Yuen MF. Relationship between HBsAg, HBcrAg 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]

87 **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.e5 [PMID: 24704525 DOI: 10.1053/j.gastro.2014.03.048]

88 **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]

89 **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: 23162861]

90 **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]

91 **Chen HP**, Shieh 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: 22773548 DOI: 10.1136/gutjnl-2011-301708]

92 **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/s12887-016-0567-0]

93 **Gevers D**, Kugathasan S, Denson LA, Vázquez-Baeza Y, Van Treuren W, Ren B, Schwager E, Knights D, Song SJ, Yassour M, Morgan XC, Kostic AD, Luo C, González A, McDonald D, Haberman Y, Walters T, Baker S, Rosh J, Stephens M, Heyman M, Markowitz J, Baldassano R, Griffiths A, Sylvester F, Mack D, Kim S, Crandall W, Hyams J, Huttenhower C, Knight R, Xavier RJ. The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host Microbe* 2014; **15**: 382-392 [PMID: 24629344 DOI: 10.1016/j.chom.2014.02.005]

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

95 **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]

96 **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]

97 **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]

98 **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: 29857150 DOI: 10.1016/j.cgh.2018.05.024]

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

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

101 **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]

102 **Ji J**, Sundquist 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]

Baishideng®

Published By Baishideng Publishing Group Inc
7041 Koll Center Parkway, Suite 160, Pleasanton, CA 94566, USA
Telephone: +1-925-2238242
Fax: +1-925-2238243
E-mail: bpgoffice@wjgnet.com
Help Desk:http://www.f6publishing.com/helpdesk
http://www.wjgnet.com