Cohort profile: ESPRESSO (Epidemiology Strengthened by histoPathology Reports in Sweden)

This article was published in the following Dove Medical Press journal: Clinical Epidemiology

Abstract: The ESPRESSO study constitutes a novel approach to examine the etiology and prognosis of gastrointestinal disease in which histopathology plays a prominent role. Between 2015 and 2017, all pathology departments (n=28) in Sweden were contacted and asked to procure histopathology record data from the gastrointestinal tract (pharynx to anus), liver, gallbladder, and pancreas. For each individual, local histopathology IT personnel retrieved data on personal identity number, date of histopathology, topography (where the biopsy is taken), morphology (biopsy appearance), and where available free text. In total, between 1965 and 2017, histopathology record data were available in 2.1 million unique individuals, but the number of data entries was 6.1 million because more than one biopsy was performed in many of the study participants. Index individuals with histopathology data were matched with up to five controls from the general population. We also identified all first-degree relatives (parents, children, full siblings), and the index individual’s first spouse. The total study population consisted of 13.0 million individuals. Data from all the study participants have been linked to Swedish National Healthcare Registers allowing research not only on such aspects as fetal and perinatal conditions and the risk of future gastrointestinal disease but also on the risk of comorbidity and complications (including cancer and death). Furthermore, the ESPRESSO database allows researchers and practitioners to identify diagnoses and disease phenotypes not currently indexed in national registers (including disease precursors). The ESPRESSO database increases the sensitivity and specificity of already-recorded diseases in the national health registers. This paper is an overview of the ESPRESSO database.

Keywords: cohort, gallbladder, gut, liver, pancreas, population-based, colon, pathology, endoscopy, celiac, inflammatory bowel disease, colitis, esophagus

Introduction to the ESPRESSO cohort

Sweden has a population of about 10 million inhabitants,¹ with a total life expectancy of 82.4 years (males 80.7 years, females 84.0 years).² Healthcare delivery is funded by a decentralized, taxpayer-funded system. In 2014, healthcare expenditures comprised 11.9% of the country’s gross domestic product.³ Healthcare use is monitored by “Statistics Sweden” (Swedish: Statistiska Centralbyrån, SCB), and healthcare registers have been the source of invaluable population-based registry linkage research.⁴ Despite the importance of histopathology data in routine healthcare, none of the Swedish government-administered national healthcare registers (with the exception of limited cancer data in the “Swedish Cancer Register”⁵) contains any histopathology information.
From 2004 to 2007, the principal investigator (JFL) of the ESPRESSO database published several scientific papers on coeliac disease\(^6\)\(^-\)\(^8\) using relevant International Classification of Disease (ICD) codes in the Swedish Inpatient Register.\(^9\) However, it soon became evident that using inpatient data for the identification of coeliac disease jeopardized overestimating the risk of complications given that coeliac patients admitted to hospital represent individuals with a more severe disease (selection bias)\(^10\) than the average patient.\(^11\),\(^12\)

We therefore hypothesized that small intestinal biopsy record data might help to identify individuals with coeliac disease, and furthermore, are more representative of all patients with coeliac disease. Between 2006 and 2008, all Swedish pathology departments (n=28) were contacted and consented to export data for a histopathology-based coeliac database. An important characteristic of this coeliac database is that it contains information on inflammation and “normal mucosa”\(^13\),\(^14\) – phenotypes that lack ICD codes and therefore cannot be identified through national healthcare registers – but the database also enabled us to identify patients with coeliac disease who healed over time (mucosal healing)\(^15\),\(^16\) (first biopsy with villous atrophy, but the second biopsy showed no persistent villous atrophy).

It has become clear over the years that research on other gastrointestinal diseases (eg, microscopic colitis and fatty liver disease) may also benefit from histopathology data. Accordingly, between October 12, 2015, and April 10, 2017, we collected all gastrointestinal histopathology report data from the 28 Swedish pathology departments.

Data collection

In Sweden, gastrointestinal biopsies and surgical specimens are classified according to the SnoMed classification system. SnoMed (or SnoMed CT) is a multilingual clinical healthcare terminology (http://www.ihtsdo.org/snomed-ct/) that was jointly developed by the National Health Service (NHS) in England and the College of American Pathologists.

To minimize heterogeneity due to different reporting from individual pathology departments and to ensure that standard information was delivered from each centre, JFL in 2015 contacted the two companies that deliver the technical systems needed for SnoMed recording and storage in Sweden, namely “CompuGroup Medical” (system platform: “Analytix”) and “Tieto-Enator” (system platform: “Sympathy”). Together with these companies, JFL constructed a dedicated structured query language (SQL) question, which was then delivered as part of an update of the Analytix and Sympathy systems. This SQL question enabled local users to perform the search automatically, with a predefined search algorithm (roughly “the gastrointestinal tract including the liver, gallbladder, and pancreas” up until the date of search).

The algorithm was distributed in 2015, but because of local requirements (including a specific request for locally tailored research applications), the search was delayed and performed in most pathology departments in 2016 (up until April 10, 2017). All Swedish pathology departments contributed data (percentages of all biopsies from Sweden are reported within brackets): Borås (contributed 1.95% of all gastrointestinal histopathology reports in Sweden), Eskilstuna (1.56%), Falun (2.96%), Gävle (2.14%), Halmstad (2.03%), Jönköping (3.28%), Kalmar (2.30%), Karlskrona (1.42%), Karlstad (3.31%), Medilab (Stockholm) (11.43%), St Göran Hospital (1.51%), Göteborg (6.61%), Lund-Malmö-Helsingborg-Kristianstad (14.14%), Skövde (3.27%), Stockholm (12.89%), Sunderbyrn (1.98%), Sundsvall (1.65%), Trollhättan (2.69%), Umeå (4.57%), Uppsala (4.14%), Västerås (1.96%), Växjö (2.07%), Örebro (6.37%), and Linköping-Norrköping (3.77%).

Local IT technicians performed searches and saved data on 1) date of biopsy, 2) personal identity number (PIN),\(^4\) 3) morphology, 4) topography, and 5) free text of the histopathology report (available from ten departments: Eskilstuna, Falun, Jönköping, Karlskrona, Medilab [Stockholm], St Göran [Stockholm], Skövde, Sunderbyrn, Västerås, and Örebro). Computerized search results were delivered to the researchers. ML then cleaned the data and merged all the files. It should be noted that the histological classification of the ESPRESSO database is based on information recorded in histopathology reports. We did not request any tissue collection or DNA and were unable to collect histopathology image data.

Eligibility criteria

We requisitioned all electronic histopathology reports in Sweden with a topography code from T56 to T69 corresponding to the liver, gallbladder, pancreas, and the bowel (from the pharynx to the anus; Figure 1). A small number of histopathology reports occurred before 1965 (n=13) and were omitted from the cohort for fear of misclassification. In total, we identified 2,109,579 unique individuals with a histopathology report between 1965 and April 2017. The number of data sets contained more data entries (6.1 million) than the number of individuals because many of the individuals (53.8%) had been biopsied more than once.
Figure 2 shows the number of histopathology reports collected in the ESPRESSO database per year. The number of histopathology reports increased gradually until 2015. The lower number of reports in 2016 and 2017 (for which we have data up until April) reflects that the first centers sent their data to us (the researchers) in the late 2015 but not thereafter.

Each (index) individual with a histopathology report was matched with up to five controls for age, sex, calendar year (of biopsy), and county of residence from the Total Population Register. We also identified all first-degree relatives of each index individual and his or her controls (parents, children, full siblings) as well as first spouse. First-degree relatives and spouses may or may not have had a biopsy.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Liver (T56)</td>
<td>349,794</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gallbladder (T57)</td>
<td>253,628</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bile tract (T58)</td>
<td>35,323</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pancreas (T59)</td>
<td>61,542</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pharynx (T60)</td>
<td>11,499</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tonsils* (T61)</td>
<td>42,365</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Esophagus (T62)</td>
<td>360,119</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stomach (T63)</td>
<td>1,270,372</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Duodenum (T64)</td>
<td>701,290</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jejunum–ileum (T65)</td>
<td>328,129</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Appendix (T66)</td>
<td>200,914</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Colon (T67)</td>
<td>1,618,953</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rectum (T68)</td>
<td>771,511</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anus (T69)</td>
<td>91,807</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 1 Number of gastrointestinal histopathology reports from 1965 to April 2017 in Sweden. Notes: The number of unique individuals is smaller than the number of entries. *Includes adenoid tissue. †T62 (for ileum) allows for differentiation between the jejunum and ileum. ‡T67 can be further divided into caecum, pars ascendens, right flexure, transverse, left flexure, descendens, and the sigmoid.
The total study population contained 12,983,573 individuals. The study population exceeded the total Swedish population (roughly 10 million people) because >10 million unique individuals had lived in Sweden during the years when histopathology data were collected (Figure 2).

Through linkage to the Total Population Register with the date of death and emigration, it is possible to follow individuals over time. The median follow-up from the first biopsy date until 2017 was 12 years (we had data on the date of death and emigration until December 31, 2017). Males constituted 45.8% of the individuals having a histopathology record and the median age at first biopsy was 58 years.

Given the nature of the study and the almost 100% follow-up data of individuals (tracked through their PIN), we expect virtually no loss to follow-up. Concerning the histopathological follow-up, most patients had one histopathology report (median =1). Participants were never contacted directly nor did they receive any questionnaires or undergo any clinical assessments as part of the ESPRESSO database. Instead, all data originated from the histopathology reports and from linked national Swedish registers (Figure 3).

**Linkage to background data**

For each study participant (n=13.0 million), Statistics Sweden delivered information on vitality status (dates of birth and death) and immigration/emigration. Statistics Sweden also supplied data on sex, age, county of residence, civil status, income, education, number of children, occupation, socioeconomic status, and nationality. Most of the background data originated from the LISA (longitudinal integrated database for health insurance and labor market studies) database.

**Linkage to other national registers**

Health data (serving as exposures, outcome measures, and covariates) were obtained from the Swedish national registers

---

**Figure 3** Main linkages in the ESPRESSO database.

**Note:** Pseudonymized data, and the explanation is given in the “Ethics” section.

**Abbreviations:** PIN, personal identity number; GI, gastrointestinal.
mainly by the National Board of Health and Welfare (Swedish: Socialstyrelsen). Approximately 97% of the Swedish healthcare system is under public authority, permitting easy and inexpensive access to comprehensive registry data. For each individual, we obtained data from the “Causes of Death Register” (covering >99% of all deaths),18 “the Swedish Cancer Register” (began in 1958, where >96% of all malignancies are recorded), the “Patient Register” (began in 1964, with hospital-based outpatient data since 2001; the positive predictive value (PPV) of the Patient Register is usually –90%),9 the “Medical Birth Register” (antenatal and perinatal data on >98% of all births in Sweden since 1973 and includes data on early pregnancy smoking since 1982 and early pregnancy body mass index since 1992), and the “Swedish Prescribed Drug Register” (established in July 2005).19 Additional linkage, including the Swedish Twin Register, is expected in the future.20 Figure 3 shows the main linkages of the ESPRESSO study.

Validation of histopathology codes
In 2009, we validated small intestinal histopathology data for coeliac disease against patient chart data.21 This chart validation found that 95% (108/114) of the individuals with small intestinal villous atrophy had a clinical diagnosis of coeliac disease (PPV, 95%). This figure is higher than when receiving a physician-assigned diagnosis of coeliac disease in Sweden (86%).22 Patient chart reviews demonstrated that 79% of the individuals with villous atrophy had gastrointestinal symptoms before biopsy and that the 88% with villous atrophy had positive coeliac disease serology before the first biopsy, further confirming that individuals with villous atrophy were likely to suffer from coeliac disease. The coeliac disease validation also contained individuals with small intestinal inflammation but no villous atrophy (n=39). We reviewed symptoms and signs of these individuals.21 In a related validation, we were able to examine patient charts of 112 individuals with normal mucosa but positive coeliac antibodies.21

In a third validation, we examined patient charts from individuals with a histopathology report indicating microscopic colitis.24 In total, 200/211 patients with a histopathology diagnosis of microscopic colitis were confirmed as also having a clinical diagnosis of microscopic colitis after chart review, yielding a PPV of 95% (95% CI =91%–97%). The most common symptoms in patients with microscopic colitis were diarrhea (seen in 96%), weight loss (24%), and abdominal pain (13%).

Relationship to IBD
Histopathology use can serve as a means to identify patients with inflammatory bowel disease (IBD) (Figure 4). Up until 2001, only inpatient diagnoses of IBD were registered in the Swedish Patient Register;9 even after that, only patients admitted to hospitals are recorded. Gastrointestinal histopathology data of Swedish IBD patients2,25,26 have revealed that such histopathology data can be used to ascertain the actual date of IBD diagnosis. Figure 4 illustrates the different incidence rates for IBD in Sweden using different definitions, where we planned to use “≥2 relevant ICD codes for IBD or one relevant SnoMed code and one relevant ICD code for IBD” (green line) as opposed to “only ≥2 relevant ICD codes for IBD” since the latter may underestimated incidence before the introduction of outpatient data in 2001. Serial biopsy data can also be used to evaluate mucosal healing in IBD. Additional data on validation against IBD can be found in the Supplementary Material.

Linkage to procedure codes in the Patient Register also allows the ESPRESSO database to add information on endoscopy (Esophagogastroduodenoscopy: 2861, 2880, 2881, 4480, 4483, 4486, 4487, 4488, 4489, 4490, 4686, 4687, 9003, 9004, 9021, UJC, UJD, UJF02, UJF05. Colonoscopy and sigmoidoscopy: 9011, 9012*, 9023, 4685*, 4688, 4689, 4674, 4684, UJF32, UJF35, UJF42*, UJF45* (sigmoidoscopy codes are marked with *). Endoscopic retrograde cholangiopancreatography: 9014, 5388, 5394, UJK02, UJK05).

Morphology data
The Swedish version of the SnoMed system allows users to assign morphology (M codes) and diagnostic codes (D codes). A diagnostic code is regarded as a kind of morphology code. A selection of diagnostic codes relevant to the ESPRESSO database is listed in Table 1.

M-codes range from normal (macroscopically: M00100, microscopically: M00110, chromosomally normal: M00150) to minimal abnormality (M01100) and to more severe abnormalities. A few M-codes are organ specific (eg, ectopic pancreas, M26020 but also ectopic bowel and stomach tissue), ulcers (M38000 and then specifically peptic ulcer M38090), chronic persisting hepatitis (M43001), chronic autoimmune hepatitis (M43005), and reactive gastritis (M69400). M-codes may also signal the existence of specific pathogens (eg Helicobacter [ME1370]) and Giardia lamblia (ME4416), but can also be used for cancer staging.
In a clinical setting, most M-codes must be combined with a topography code to make sense that a patient has a morphology code suggesting that unspecific inflammation is difficult to interpret without knowing if the inflammation occurs in the kidney, liver, or skin. M40-M43, M44000, and M470-471 are all different forms of inflammation, sometimes with additional information (when seen in the colorectal part: collagenous colitis M40600 and lymphocytic inflammation M4717 represent microscopic colitis). End-stage morphology, such as fibrosis (M49000-04) and cirrhosis (M495, with subheadings for primary biliary cirrhosis), fat infiltration (M50080 and M55200), necrosis (M54 and necrotic fat tissue M54110), infarction (M547), and atrophy (M58) can also be registered.

The SnoMed system makes it possible for pathologists to classify degree of atypia (M697), squamous cell and glandular cell atypia (with subheadings for the degree of atypia), hyperplasia (M720), and then specific M-codes for adenomatous hyperplasia (M7242), metaplasia (M73), and dysplasia (M74), polyps (M76800), and in combination with other codes (fibrous dysplasia, M74910). Furthermore, there are M-codes for benign tumors (M800), suspected cancer (the code M801 also includes in situ cancer), and metastases (M80106). Other M80-M99 codes represent different forms of cancer (eg, hepatocellular cancer, M817, with several subheadings; hepatoblastoma, M897; carcinoids, M82401; and familial polyposis coli, M82200). Most of the M80-99 codes refer to cancers outside the gastrointestinal tract.

The SnoMed system makes it possible for pathologists to classify degree of atypia (M697), squamous cell and glandular cell atypia (with subheadings for the degree of atypia), hyperplasia (M720), and then specific M-codes for adenomatous hyperplasia (M7242), metaplasia (M73), and dysplasia (M74), polyps (M76800), and in combination with other codes (fibrous dysplasia, M74910). Furthermore, there are M-codes for benign tumors (M800), suspected cancer (the code M801 also includes in situ cancer), and metastases (M80106). Other M80-M99 codes represent different forms of cancer (eg, hepatocellular cancer, M817, with several subheadings; hepatoblastoma, M897; carcinoids, M82401; and familial polyposis coli, M82200). Most of the M80-99 codes refer to cancers outside the gastrointestinal tract.

The SnoMed system makes it possible for pathologists to classify degree of atypia (M697), squamous cell and glandular cell atypia (with subheadings for the degree of atypia), hyperplasia (M720), and then specific M-codes for adenomatous hyperplasia (M7242), metaplasia (M73), and dysplasia (M74), polyps (M76800), and in combination with other codes (fibrous dysplasia, M74910). Furthermore, there are M-codes for benign tumors (M800), suspected cancer (the code M801 also includes in situ cancer), and metastases (M80106). Other M80-M99 codes represent different forms of cancer (eg, hepatocellular cancer, M817, with several subheadings; hepatoblastoma, M897; carcinoids, M82401; and familial polyposis coli, M82200). Most of the M80-99 codes refer to cancers outside the gastrointestinal tract.

The SnoMed system makes it possible for pathologists to classify degree of atypia (M697), squamous cell and glandular cell atypia (with subheadings for the degree of atypia), hyperplasia (M720), and then specific M-codes for adenomatous hyperplasia (M7242), metaplasia (M73), and dysplasia (M74), polyps (M76800), and in combination with other codes (fibrous dysplasia, M74910). Furthermore, there are M-codes for benign tumors (M800), suspected cancer (the code M801 also includes in situ cancer), and metastases (M80106). Other M80-M99 codes represent different forms of cancer (eg, hepatocellular cancer, M817, with several subheadings; hepatoblastoma, M897; carcinoids, M82401; and familial polyposis coli, M82200). Most of the M80-99 codes refer to cancers outside the gastrointestinal tract.

The SnoMed system makes it possible for pathologists to classify degree of atypia (M697), squamous cell and glandular cell atypia (with subheadings for the degree of atypia), hyperplasia (M720), and then specific M-codes for adenomatous hyperplasia (M7242), metaplasia (M73), and dysplasia (M74), polyps (M76800), and in combination with other codes (fibrous dysplasia, M74910). Furthermore, there are M-codes for benign tumors (M800), suspected cancer (the code M801 also includes in situ cancer), and metastases (M80106). Other M80-M99 codes represent different forms of cancer (eg, hepatocellular cancer, M817, with several subheadings; hepatoblastoma, M897; carcinoids, M82401; and familial polyposis coli, M82200). Most of the M80-99 codes refer to cancers outside the gastrointestinal tract.

The SnoMed system makes it possible for pathologists to classify degree of atypia (M697), squamous cell and glandular cell atypia (with subheadings for the degree of atypia), hyperplasia (M720), and then specific M-codes for adenomatous hyperplasia (M7242), metaplasia (M73), and dysplasia (M74), polyps (M76800), and in combination with other codes (fibrous dysplasia, M74910). Furthermore, there are M-codes for benign tumors (M800), suspected cancer (the code M801 also includes in situ cancer), and metastases (M80106). Other M80-M99 codes represent different forms of cancer (eg, hepatocellular cancer, M817, with several subheadings; hepatoblastoma, M897; carcinoids, M82401; and familial polyposis coli, M82200). Most of the M80-99 codes refer to cancers outside the gastrointestinal tract.

The SnoMed system makes it possible for pathologists to classify degree of atypia (M697), squamous cell and glandular cell atypia (with subheadings for the degree of atypia), hyperplasia (M720), and then specific M-codes for adenomatous hyperplasia (M7242), metaplasia (M73), and dysplasia (M74), polyps (M76800), and in combination with other codes (fibrous dysplasia, M74910). Furthermore, there are M-codes for benign tumors (M800), suspected cancer (the code M801 also includes in situ cancer), and metastases (M80106). Other M80-M99 codes represent different forms of cancer (eg, hepatocellular cancer, M817, with several subheadings; hepatoblastoma, M897; carcinoids, M82401; and familial polyposis coli, M82200). Most of the M80-99 codes refer to cancers outside the gastrointestinal tract.

The SnoMed system makes it possible for pathologists to classify degree of atypia (M697), squamous cell and glandular cell atypia (with subheadings for the degree of atypia), hyperplasia (M720), and then specific M-codes for adenomatous hyperplasia (M7242), metaplasia (M73), and dysplasia (M74), polyps (M76800), and in combination with other codes (fibrous dysplasia, M74910). Furthermore, there are M-codes for benign tumors (M800), suspected cancer (the code M801 also includes in situ cancer), and metastases (M80106). Other M80-M99 codes represent different forms of cancer (eg, hepatocellular cancer, M817, with several subheadings; hepatoblastoma, M897; carcinoids, M82401; and familial polyposis coli, M82200). Most of the M80-99 codes refer to cancers outside the gastrointestinal tract.

In a clinical setting, most M-codes must be combined with a topography code to make sense that a patient has a morphology code suggesting that unspecific inflammation is difficult to interpret without knowing if the inflammation occurs in the kidney, liver, or skin. M40-M43, M44000, and M470-471 are all different forms of inflammation, sometimes with additional information (when seen in the colorectal part: collagenous colitis M40600 and lymphocytic inflammation M4717 represent microscopic colitis). End-stage morphology, such as fibrosis (M49000-04) and cirrhosis (M495, with subheadings for primary biliary cirrhosis), fat infiltration (M50080 and M55200), necrosis (M54 and necrotic fat tissue M54110), infarction (M547), and atrophy (M58) can also be registered.

The SnoMed system makes it possible for pathologists to classify degree of atypia (M697), squamous cell and glandular cell atypia (with subheadings for the degree of atypia), hyperplasia (M720), and then specific M-codes for adenomatous hyperplasia (M7242), metaplasia (M73), and dysplasia (M74), polyps (M76800), and in combination with other codes (fibrous dysplasia, M74910). Furthermore, there are M-codes for benign tumors (M800), suspected cancer (the code M801 also includes in situ cancer), and metastases (M80106). Other M80-M99 codes represent different forms of cancer (eg, hepatocellular cancer, M817, with several subheadings; hepatoblastoma, M897; carcinoids, M82401; and familial polyposis coli, M82200). Most of the M80-99 codes refer to cancers outside the gastrointestinal tract.

The SnoMed system makes it possible for pathologists to classify degree of atypia (M697), squamous cell and glandular cell atypia (with subheadings for the degree of atypia), hyperplasia (M720), and then specific M-codes for adenomatous hyperplasia (M7242), metaplasia (M73), and dysplasia (M74), polyps (M76800), and in combination with other codes (fibrous dysplasia, M74910). Furthermore, there are M-codes for benign tumors (M800), suspected cancer (the code M801 also includes in situ cancer), and metastases (M80106). Other M80-M99 codes represent different forms of cancer (eg, hepatocellular cancer, M817, with several subheadings; hepatoblastoma, M897; carcinoids, M82401; and familial polyposis coli, M82200). Most of the M80-99 codes refer to cancers outside the gastrointestinal tract.
increase sensitivity and specificity in identifying disease (eg, fatty liver disease), identify disease precursors (early stages of dysplasia), and grade the severity of disease and disease extent (using topography codes to define the location of disease, eg, transverse colon, T674). Multiple longitudinal biopsies allow researchers to follow the mucosal response to treatment over time (eg, repeated biopsies in IBD patients who have started on biological treatment).

### Free text

Certain histopathology reports contain free text. This text can be scanned manually or by computerized algorithms to identify diagnoses or to characterize patients and their lesions. Table 3 exemplifies reports of free text.

### Ethics

Preceding data collection, the general outline of the ESPRESSO study was approved by the Stockholm Ethics

---

**Table 2** Comparison between Swedish hospital-based data and the ESPRESSO histopathology data

<table>
<thead>
<tr>
<th>Identification through hospital-based data</th>
<th>Identification through histopathology registers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hospital admission may vary with age (especially in children and old people) and disease activity</td>
<td>When a diagnosis is conditional on histopathology, then, histopathology data represent average individuals with GI disease, irrespective of their age, sex, and disease activity, thereby minimizing selection bias</td>
</tr>
<tr>
<td>Very low sensitivity for conditions that lack ICD codes (eg, microscopic colitis) or when such codes were introduced late into the ICD system (eg, eosinophilic esophagitis)</td>
<td>Applying histopathology data is the only way to identify several diseases (eg, microscopic colitis and serrated adenomas) and has been the gold standard for coeliac disease in children and remains so in adults</td>
</tr>
<tr>
<td>The date of hospital admission may not correlate with the date of diagnosis. The first inpatient diagnosis of GI disease is sometimes registered only years later</td>
<td>The date of first histopathology usually correlates with the actual diagnostic date. This information is of particular importance when examining complications in, and risk factors for, diseases (eg, IBD, Figure 4)</td>
</tr>
</tbody>
</table>

---

**Table 3** Examples of free text in the ESPRESSO database

**M-code: M72040: hyperplastic polyp**

Biopsy no. 3 shows a hyperplastic polyp. Biopsy no 7 demonstrates active ulcerative colitis with severely aberrant crypt pattern, intense inflammation and crypt abscesses, and partly the surface epithelium cannot be seen. The findings are consistent with an ulcerative proctitis. Biopsy I shows a normal ileum, biopsies II–VI. Normal colon. III. Hyperplastic polyp in the colon. VII. Active ulcerative colitis

**M-code: M49590: primary biliary cirrhosis**

The biopsy shows microscopic liver tissue with a preserved basic structure. Hepatocytes are characterized by mild anisokaryosis. The sinusoids are partly dilated. There are only a small number of inflammatory cells in the lobuli. Ordinary Kupffer cell activity. Iron staining is negative. In the portal areas, there is a small increased amount of connective tissue. There are areas with a somewhat reduced number of biliary ducts. Preserved bile ducts show ordinary epithelium. There is a mild-to-moderate increase in the number of inflammatory cells in the portal areas. No granulomas are observed. The border between lobuli and the portal zones is somewhat irregular. The biopsy is consistent with primary biliary cirrhosis (stages I–II)

**M-code: M47170: microscopic colitis (lymphocytic colitis)**

In the biopsies from the colonic mucosa is seen a chronic inflammation with an increased number of lymphocytes. I see no signs of collagenous colitis, nor is there any atypia or malignancy. Instead, the biopsy is consistent with microscopic colitis

**M-code: M00110: microscopically normal mucosa in the duodenum**

In both specimens from the duodenal mucosa, the architecture is normal with long and slender villi. The villous epithelium is normal, and there is no increase in the number of intraepithelial lymphocytes. There is nothing pathological in the deeper parts of the mucosa. Histologically, this is a normal duodenal mucosa, strongly arguing against coeliac disease

**M-code: M74009: severe dysplasia in the (pyloric) antrum**

In some areas, there is a substantially normal stomach mucosa of antrum type, but in most parts of the mucosa there is a strong atypia. The foveole are irregular, and the cylinder epithelium is severely atypical. I cannot confirm that there is invasive growth. In some areas, the atypical mucosa has eroded. In all biopsies, there is hence severe dysplasia. Invasiveness cannot be confirmed, but this is clearly a precancerous lesion in the antral mucosa

**M-code: D6214: indeterminate colitis (IBD-U)**

The biopsies consist of separate colorectal biopsies. Microscopically, there is colorectal mucosa with a rich number of inflammatory cells in the lamina propria, among which granulocytes can be observed. There are an increased number of intraepithelial lymphocytes and granulocytes. In addition, the gland architecture is aberrant. No ulcerations. No granulomas or crypt abscesses. No atypia or malignancy. The specimen is suggestive of IBD, but there are no convincing signs of activity. In conclusion, the findings are consistent with IBD
Review Board (No. 2014/1287-31/4) on August 27, 2014. Ethics amendments have since allowed us to add data from the Prescribed Drug Register, data on health economics, and data on country of birth and nationality and to validate a number of specific diagnoses (No. 2017/1497-32; July 19, 2017). Linkage to the Swedish Twin Register\textsuperscript{20} was approved on October 30, 2017 (No. 2017/2087). A more detailed background of the study was submitted to the Stockholm Ethics Review Board (No. 2018/972-32) and approved on May 14, 2018.

The large-scale register-based nature of the ESPRESSO database determines that participants will not be directly contacted by the researchers. Therefore, the ethics review board has waived informed consent.\textsuperscript{29} To protect the integrity of the data, linked data are “pseudonymized before delivery”. This pseudonymization procedure implies that the PINs have been replaced by serial numbers by the government agencies delivering data to the researchers. However, we (researchers) have requested that a key between the PIN and the serial number should be saved at Statistics Sweden until March 2021. Such a key allows for additional linkages (eg, if some data are found to be incorrect) or updates (after relevant ethics approval). Furthermore, requesting informed consent for 13 million people would make this research not only impossible to carry out (for the economic resources needed to contact everyone and because a number of participants have deceased) but also because selective opt out\textsuperscript{30–32} would damage the validity and statistical power of the study. Vulnerable participants (such as minorities and children) would be at greater risk of being excluded. Finally, the ESPRESSO study aims to examine mortality, and because the data are retrospective, it is not possible to go back in time to approve study participation. Through the ESPRESSO database, we can avoid selection bias and have access to general population-based data. It will also allow us to examine side effects of hazardous exposures that otherwise would have been unethical to investigate in a randomized clinical trial.

The main risk with register-based research is if someone, by mistake or purposely, reveals the identity of the study participants. That is possible either through stealing the key from the National Board of Health and Welfare or through “backtracking”. Assume, for instance, that Patient X is a woman, who was born in 1977, immigrated to Sweden in 1987, is of Jamaican heritage, and is now living in one of the Northern provinces of Sweden. These bits of information could be utilized to deduce the true identity of Patient X, a deduction which is illegal. To safeguard against this potentiality, the National Board of Health and Welfare and Statistics Sweden have established a protocol in which most often only the year and month of birth (but not day) are supplied for individuals >2 years of age at first biopsy and where nationalities are aggregated into larger groups.

Ethics approvals in Sweden are not time limited, and hence no annual approval is needed.\textsuperscript{33} The ESPRESSO database has been registered with the Karolinska Institutet personal data officer (Swedish: Personuppgiftsombud).

**Potential research use of the ESPRESSO database**

As noted by the ESPRESSO acronym, the strength of this study lies in its ability to combine histopathology data with the rich resources of the Swedish National Healthcare Registers. Although it is particularly relevant to examine gastrointestinal disease that is identified in the Patient Register (to determine the phenotype of patients or to increase the validity of a diagnosis), mucosal histopathology can serve other purposes. For instance, ESPRESSO helps in the identification of patients with irritable bowel syndrome (IBS) or IBD with a relevant ICD code in the 13 million person cohort but with a possibility to stratify for mucosal appearance (normal, inflammation, other appearance, etc).

We illustrate how gastrointestinal histopathology data can improve epidemiological research in the following sections.

**Infections**

Through linkage to the Patient Register, it is possible to examine gastrointestinal diseases and sepsis.\textsuperscript{34} Infections may potentially cause gastrointestinal disease (eg, multiple reports indicate an association between rotavirus and coeliac disease),\textsuperscript{35,36} but gastrointestinal disease per se may also predispose to infections.\textsuperscript{37} Other times, infections may mimic gastrointestinal disease (intestinal tuberculosis can be misjudged as Crohn’s disease, tuberculosis as coeliac disease, etc). It is therefore important to combine histopathology data and data from the Patient Register and the Cancer Register (as well as other registers).

**Cancer**

Several gastrointestinal diseases are characterized by inflammation, and inflammatory activity has been linked to a number of cancers.\textsuperscript{16,38–47} Earlier studies suggested that Barrett’s esophagus may be a risk factor for ear–nose–throat cancer. Although coeliac disease has been linked to a lower risk of lung cancer\textsuperscript{48} (possibly because of lower rates of smoking among patients),\textsuperscript{49} these same patients are probably at increased risk of both liver cancer and IBD.\textsuperscript{50,51} IBD of...
itself has been associated with thymic cancer. Inflammation has also been linked to malignant melanoma and to nonmelanoma skin cancer. The linkage to the Prescribed Drug Register allows researchers (and practitioners) to add data on medication and thereby to disentangle the effects of inflammation and treatment (sometimes for inflammation).

**Psychiatric disease**

In recent years, interest in the brain–gut axis has increased dramatically. In addition, research has increasingly focused on the potential implications of antibiotic use, the microbiota, and incident gastrointestinal disorder in psychiatric disease. Our research group has previously demonstrated an association between coeliac disease and depression (and also other psychiatric disorders and even suicide). Crohn’s disease has been linked to depression, and some data suggest that the presence and extent of depression influences the risk of complications in IBD. Compellingly, antidepressive treatment can sometimes ameliorate the symptoms and situation of patients with gastrointestinal pain.

Through linkage to normal mucosa and serology data, we have detected a potential association between early (potential) gastrointestinal disease and autism. Some patients with gastrointestinal disease describe a phenomenon known as “brain fog”, whereas others may experience ataxia. Depression is also more common in patients with colorectal cancer; however, the role of inflammation for this risk increase is unknown. Attention-deficit hyperactivity disorder is another neuropsychiatric disorder that has been linked to gastrointestinal disease, including liver disease.

**Neurology**

Neurological conditions may be congenital or acquired. Many children with early-life neurological disease suffer from concomitant gastrointestinal tract conditions. Here, ESPRESSO offers the advantage to differentiate between children with biopsy-verified abnormalities and children with a normal mucosa. Dementia has been linked to IBD and although a recent Swedish study found no association between IBD and Parkinson disease, Danish researchers just reported a positive association. Other disorders that have been linked to gastrointestinal disease include epilepsy, migraine, and neuropathy. Via the ESPRESSO database, it is possible to examine nonspecific gastrointestinal inflammation and its association with neurological disease.

**Cardiovascular disease**

The ESPRESSO database contains hundreds of thousands of individuals with chronic inflammation in the gut. Thus, researchers are able to examine the association with chronic inflammation, specific gastrointestinal disease (other than coeliac diseases), and cardiovascular disease. Inflammation may not only be a risk factor for ischemic heart disease and stroke but also for tromboembolism.

**Respiratory disease**

Respiratory disease is a major cause of death worldwide and has been linked to a number of gastrointestinal disorders. Some data indicate that patients with gastrointestinal disease do not respond to certain vaccinations; if this assertion proves to be true, such patients may be at increased risk of severe infections (respiratory infections, including influenza and respiratory syncytial virus in small children and also pneumonia). Certain inflammatory gastrointestinal conditions, such as eosinophilic esophagitis (characterized by a typical histopathology in the esophagus) may be linked to allergic disorders, including asthma.

**Musculoskeletal disease and connective tissue disorders**

Some connective tissue disorders have autoimmune traits and may be linked to gastrointestinal autoimmune disease. Because of malabsorption and calcium deficiency, bone mineral density is often decreased in patients with gastrointestinal disease. Although the ESPRESSO database does not contain data on bone mineral density, it offers an opportunity to calculate precise risk estimates for fractures (especially for osteoporotic fractures), gastrointestinal diseases, with chronic inflammation and poor nutrient uptake, may predispose to osteoporosis and fractures.

**Renal disease**

A PubMed search in May 2018 revealed 120 publications on gastrointestinal disorders and glomerulonephritis. We have previously shown that coeliac disease is associated with end-stage renal disease as well as with milder renal disease. The Patient Register contains data on procedures, implicating that data on dialysis and renal transplantation can be ascertained as per the linked ESPRESSO database.

**Pregnancy and perinatal health**

By linkage to the Medical Birth Register and the Multigeneration Register (the latter is part of the Total Population Register), it is possible to examine pregnancy outcome in offspring to parents (both mothers and fathers) with gastrointestinal disease. Many women with chronic gastrointestinal diseases are encouraged to give birth through cesarean section, which may have serious consequences for offspring health. Of note,
the ESPRESSO database includes information on procedures. Two recent papers failed to detect any substantial impact on pregnancy outcome from endoscopy or liver biopsy during pregnancy. Yet, pregnancy outcome may not only be seen as a consequence of parental gastrointestinal disease, it could also be a risk factor for future gastrointestinal disease in the offspring. Finally, certain gastrointestinal diseases have been linked to IBD; tetracycline use may induce fatty liver and gastrointestinal bleeding. Oral contraceptives have been linked to gastrointestinal disease. Certain drugs aim to change the microenvironment in the gastrointestinal tract (eg, proton pump inhibitors); others, such as antibiotics, will influence the microflora, potentially impacting the risk of gastrointestinal disease. Medication data may also influence liver pathology. Linkage to the Patient Register will also allow researchers to calculate the Charlson comorbidity index or other comorbidity indices when needed for studies in the ESPRESSO database.

The ESPRESSO database also contains limited data on skin, eye, and ear–nose–throat disease.

Use of medications
Linkage to the Prescribed Drug Register unlocks opportunities for pharmaco-epidemiological research in gastrointestinal disease. Certain drugs aim to change the microenvironment in the gastrointestinal tract (eg, proton pump inhibitors); others, such as antibiotics, will influence the microflora, potentially impacting the risk of gastrointestinal disease. Medication data can also be used to identify individuals with milder disease in which hospital admission is rarely needed (eg, hypertension or hyperlipidaemia). NSAIDs have been linked to gastric ulcers and gastrointestinal bleeding. Oral contraceptives have been linked to IBD; tetracycline use may induce fatty liver development; and some data suggest that aspirin could have an antitumor effect. In all, linkage to Prescribed Drug data enhances the quality and usefulness of histopathology data.

Discussion
In this profile paper, we outline the background, structure, and potential use of the ESPRESSO cohort. The strengths of the database include the large number of participants, the longitudinal records with histopathology, actual data on histopathology (including normal mucosa, inflammation, and cancer precursors), linkage to the Swedish National Registers, complementary medical data (medical diagnoses, medication use, etc), and data on first-degree relatives and first spouses, which could serve as important secondary controls to examine intrafamilial confounding when that is an issue. The ESPRESSO database also allows for comparisons between individuals with a certain disease, as diagnosed through histopathology (eg, microscopic colitis), and individuals with a normal mucosa on biopsy, hereby minimizing the influence of health-seeking pattern and reducing confounding due to underlying comorbidity.

The ESPRESSO cohort has some limitations. Figure 2 is likely to reflect both an increased use of histopathology and an increase in reporting of actual biopsies. The coverage of the database may be suboptimal, especially before the year 1990. We do not have access to biological data. This limitation prohibits DNA analyses, or any other laboratory analysis, if the ESPRESSO data are not linked to other data sources containing relevant biological data. That is the case with our planned linkage to the Swedish Twin Register.

We cannot rule out a certain amount of selection bias. Patient characteristics may differ from that of the general patient population, if histopathology report data are used to identify diseases that may otherwise not require biopsy, or where biopsy is typically carried out only for advanced disease. The ESPRESSO database is likely to have lower coverage for certain gastrointestinal diseases in patients where endoscopy and, eg, liver biopsy are sometimes avoided due to other concurrent conditions (eg, pregnancy). The new recommendations allowing for a non-biopsy celiac diagnosis in certain children, but not in adults, mean that not all children with incident celiac disease after 2012 may have been identified. It is also possible that in a population where cancer is more common (older people), endoscopies are more often carried out to make sure that a gastrointestinal symptom is not caused by malignancy, and this may lead to higher number of biopsies with both normal and abnormal findings among the elderly. In addition, reference individuals may differ with regard to, eg, age and sex, as they were matched on these factors with individuals having a histopathology report.

In conclusion, the ESPRESSO database offers a tantalizing opportunity to strengthen epidemiology through histopathology.

Abbreviations
ESPRESSO, Epidemiology Strengthened by histoPathology Reports in Sweden; GI, gastrointestinal; IBD, inflammatory
bowl disease; IBD-U, unclassified IBD; PIN, personal identity number; PPV, positive predictive value; SnoMed CT, Systematized Nomenclature of Medicine-Clinical Terms.

**Disclosure**

Dr JFL coordinates a study on behalf of the Swedish IBD quality register (SWIBREG). This study has received funding from Janssen corporation. The authors report no other conflicts of interest in this work.

**References**


87. Tjernberg AR, Ludvigsson JF. Children with celiac disease are more likely to have attended hospital for prior respiratory syncytial virus infection. *Dig Dis Sci*. 2014;59(7):1502–1508.


Supplementary material

Additional data on linkage between the ESPRESSO (Epidemiology Strengthened by histoPathology Reports in Sweden) database and inflammatory bowel disease (IBD) in the Swedish Patient Register

The number of histopathology reports in IBD patients diagnosed since 2002 was 4.9 in Crohn’s disease, 4.8 in ulcerative colitis (UC), and 2.7 in unclassified IBD (IBD-U). Patients with Crohn’s disease with an ICD code indicative of ileal disease (L1) more often had a record of ileal biopsies in ESPRESSO than had Crohn’s disease with other locations. Crohn’s disease patients with a perianal modifier were more likely to have a biopsy from the anorectal area than patients without this modifier. Looking specifically at the proportion of IBD patients with ≥1 histopathology report, normal mucosa was seen in only 8% of Crohn’s disease patients, 6% of UC patients, and 12% of IBD-U patients, indicating that most IBD patients have abnormal mucosa.

Because histopathology data also include specimens from surgery, we examined all bowel surgeries in Swedish IBD patients. We found that the proportion of patients with a histopathology record after surgery (n=27,933) was 86% in Crohn’s disease, 83% in UC, and 94% in IBD-U. The higher proportion among patients with IBD-U may be because this is a more recent diagnosis and surgeries for IBD-U are likely to have taken place when the coverage of the ESPRESSO database was higher.