THIN Database Description

Background

- The Health Improvement Network (THIN) is collaboration between two companies: EPIC (provider of the primary care patient data that is used for medical research) and In Practice Systems (InPS) (developer and supplier of the computer software Vision used by general practitioners in the UK).
- THIN consists of the anonymised patient records that are downloaded directly from the GP's office providing timely medical information for research purposes.
- THIN data are collected in a non-intrusive fashion during the routine practice of the general practitioner (GP) and therefore reflect "real life", allowing rapid analyses using any study design of diseases and prescribed medicines and treatments.

Population Characteristics

- The patient population represented in THIN is broadly representative of the UK population.
- The THIN database contains data from 464 GP practices and over 7.5 million patients in the United Kingdom, of which more than 3 million are actively registered with the practices. The remaining patients have historical data but have either left the practice or died.
- A portion of patients in THIN overlap with those patients in GPRD™, although no method is currently validated to make this direct link.

Data Structure

- The data are organized in files by individual practice, and provide a longitudinal medical record for each patient.
- GPs use the computer software to enter data into different files:
    - **Demographic File** with information on age, sex, registration date when entering the practice, and date when leaving the practice.
    - **Medical File** includes medical diagnoses that are part of routine care or resulting from hospitalizations, consultations, or emergency care, along with the date and location (e.g., GP's office, hospital, consultant) of the event and an option for adding free text; referrals to hospitals and specialists. Free text are not included in Penn's version of the data but anonymized text can be obtained from EPIC.
    - **Prescriptions File** includes all prescriptions along with the date of prescription, formulation, strength, quantity, and dosing instructions, indication for treatment for all new prescriptions (inferred from cross reference to medical events on the same date), and events leading to withdrawal of a drug or treatment.
    - **Additional Health Data (AHD) File** includes vaccinations and prescription contraceptives; miscellaneous information such as smoking, height, weight, immunizations, pregnancy, birth, death, and laboratory results.
- Diagnoses are coded in Read codes, which are alphanumeric codes that group and define illnesses using a hierarchical nosologic system. The Read Codes are a very comprehensive coded clinical language developed in the UK and funded by the National Health Service. The codes include terms relating to observations (signs and symptoms), diagnoses, procedures, and laboratory and radiologic tests.
- Prescriptions are currently entered using Multilex codes issued by First Databank. Drug codes provide detailed information on the drug, dose, and route of administration.
- Patients in THIN are also issued a number that identifies people residing at the same address or not at the same address but are family members. This data field, in combination with the date of birth of children and the delivery code in the mother's record, can be linked to identify mothers and their children.
- Nearly all practices have laboratory data automatically downloaded into the electronic medical record and therefore have these data available for research purposes.
- Biosamples may be obtained for the purpose of performing large population-based pharmacogenetic epidemiology studies.

Last updated on 08/25/10.

○ Links

www.epic-uk.org

❖ THIN Research Considerations

○ Strengths of THIN Database

- THIN represents a defined population
  - ✓ allows investigators to study all patients with a given disease.
  - ✓ allows investigators to study control patients from the same source population from which those with the disease of interest are derived.
- The THIN is broadly representative of the UK population in general
  - ✓ therefore findings should generalize to the broader UK population.
  - ✓ therefore minimizes selection bias and improves the validity of epidemiologic studies.
- The well-defined population of the THIN allows investigators to study families and to link health events in mothers to outcomes in their children.
- Size of the database
  - ✓ can be used to study rare outcomes
- Longitudinal data since the early 1990s in some practices.
- Access to original medical records
  - ✓ allows investigators to have the opportunity through an intermediate to obtain anonymized copies of the patient's (non-electronic) medical record and a more detailed review of the patient's health history.
  - ✓ allows researchers to verify information captured on death certificates and letters from specialists, without breach of confidentiality.
- Access to laboratory data

○ Limitations of THIN Database

- THIN prescription data reflect what was prescribed by the PG.
  - ✓ This is removed from biologic ingestion because it may not have even been dispensed by the pharmacy.
- Incompleteness of data
  - ✓ Although the data entered into THIN is used by the GP as the patient's medical record, and therefore information generated by the general practitioner is expected to be complete, information from specialists as well as events that occur in the hospital may not be fully captured in the electronic medical record.
  - ✓ The database may not contain data on every patient characteristic or disease characteristic that may be required for a study (information on occupation, employment, and individual socioeconomic status is not available electronically).
  - ✓ Communication from specialists, discharge summaries from hospitals, and test results from pathology laboratories are often received in hard copy and must be manually entered into the computer. Since this can be a time consuming process, some practices will only enter information that will affect the care of the patient in the future. Therefore, with historic test results it is likely that only the abnormal results were entered onto the

computer for collection and inclusion into THIN. These are now more likely to be received and recorded electronically so the bias is removed in more recent data.
  - ✓ Minor medical events are more likely to be missed than medically significant diagnoses or events.
  - ✓ Information on treatments that are restricted by the National Health Service to specialist care (e.g., psoralen plus ultraviolet A therapy, cytotoxic/chemotherapy, Biologic therapy) may be particularly problematic.
  - ✓ Data on non-prescription medications and treatments given in the hospital are not readily available.
  - ✓ Data on non significant medical events and exposures to medications that occurred prior to enrollment in THIN and are no longer active clinical issues may also not be documented in the electronic medical record.
  - ✓ Data on important confounding variables (smoking, alcohol use, smoking, body weight and height) are only available for some patients. This is improving due to Quality Outcomes Framework initiatives to improve recording.
- Complexity and costs of computer hardware/software needed to work with THIN
  - ✓ The size and complexity of the THIN database requires adequate computer software and hardware, as well as experienced data managers.
  - ✓ The full THIN dataset received by CCEB from EPIC contains all information in THIN from 1985 to 2009, approximately 55 gigabytes. It is residing on a Sun Microsystems Enterprise™ server running the Solaris operating system in a Unix environment.
  - ✓ The data base is implemented using Oracle's® relational database. The database tools include Oracle's internet developer's suite of tools, such that clients within our server environment can access the THIN in real time using Oracle Discover®, as well as by using a SAS©-based interface. With these tools it has been possible for clients, using Ethernet-connected PCs, to download data and convert the data into several different database programs including Microsoft Access© and Visual Dbase© and statistical packages such as Stata®, SAS©, and SPSS®. A smaller 10% random sample is housed on a stand alone PC that can be queried using MySQL.
- Validated information
  - ✓ The validity of THIN has not yet been extensively studied.

❖ Requirements for Accessing the THIN Database

○ All Users:

- Must complete the Registration and Request for Help Form. Please contact Rita Schinnar at ritas@mail.med.upenn.edu

- The data access license from EPIC covers aspects of confidentiality. As part of our steps to assure that we preserve the confidentiality of the data, all users are required to abide by the CCEB's **Data Use and Approval Form.**

  ✓ Download the Data Use and Approval Agreement Form.
  ✓ Obtain signatures from either Dr. *Brian Strom* (824 Blockley Hall), or Dr. *James Lewis* (720 Blockley Hall), or Dr. *David Margolis* (815 Blockley Hall) at 423 Guardian Drive, Philadelphia, PA 19104-6021.
  ✓ Return the signed form to *Rita Schinnar* (address: CCEB, 807 Blockley Hall, 423 Guardian Drive Fax: 215-573-5315).

- **If using only the pre-collected THIN data:**

a) Must get scientific review and approval of the protocol from the **Penn IRB**.

  ✓ refer to: http://www.upenn.edu/regulatoryaffairs/human/ApplicationProcedures.html
  ✓ http://www.med.upenn.edu/pennmanual/sp/irb/sr.html
  ✓ http:/www.upenn.edu/regulatoryaffairs/human/FAQs.html

b) Must get scientific review and approval from **EPIC** through a Scientific Review Committee– therefore a copy of the Penn IRB approval, the protocol, and an SRC application should be sent to *Mustafa Dungarwalla* who will follow up with SRC approval. SRC forms are available below.

- Procedure guide for researchers - Jun2010_v5.pdf
- PROTOCOL proforma - NOV 2009 v3.doc
- PROTOCOL REQUIREMENTS - Jun2010_v4.pdf

  ✓ email contact: Mustafa.dungarwalla@epic-uk.org

- **If the study will need to go back to the practices** to obtain any additional data from the practices:

a) The researcher must also initially contact EPIC to discuss additional arrangements which need to be made through **EPIC. A quote will be prepared by *Anne Costello* at EPIC for the required administration and payments to General Practitioners.**

  ✓ refer to: http://www.epic-uk.org/ or the email contact to send the application: Anne.Costello@epic-uk.org

b) Must also get **Ethical Review** and approval of the protocol from **a Multi-centre Research Ethics Committee (MREC) via EPIC**

- ✓ refer to: http://www.epic-uk.org

c) Applications must go through EPIC. To get the application form and protocol guidelines refer to:

- ✓ email contact: Mustafa.dungarwalla@epic-uk.org

- General questions on data structure and preparing an analytic dataset can be directed to Kevin Haynes, Pharm.D., M.S.C.E. (khaynes@upenn.edu)

- Must review background readings:

  Strom BL. Overview of Automated Databases in Pharmacoepidemiology. Chapter in Pharmacoepidemiology (Strom BL, Ed). 4th edition. Chichester, England: John Wiley. 2005. pp.219-222.

❖ Examples of THIN Publications

○ Publications by CCEB Investigators

Lewis JD, Schinnar R, Bilker WB, Wang X, Strom BL. Validation studies of the Health Improvement Network (THIN) database for pharmacoepidemiology research. *Pharmacoepidemiology and Drug Safety. 2006 15: (S1):S233-234.*

BACKGROUND: The Health Improvement Network (THIN) is a new medical records database that contains records from general practices some of which have or continue to participate in the General Practice Research Database (GPRD™) and others that never participated in GPRD™. We sought to replicate in THIN well-established associations from the medical literature and to compare results from the GPRD™ practices to the non-GPRD™ practices within THIN. METHODS: Using THIN data from 1986-2003, we conducted case-control studies of associations between diseases (e.g., hypertension and stroke) and between diseases and drugs (e.g., aspirin and colon cancer). Conditional logistic regression was used to calculate odds ratios adjusted for potential confounders. Differences between GPRD™ and non-GPRD™ practices were assessed by testing for a statistical interaction by practice type in each outcome-exposure association. RESULTS: We observed the expected positive associations ($p < 0.05$) of stroke with hypertension and diabetes mellitus; of myocardial infarction with hypertension, hypercholesterolemia, obesity, and smoking; and of peptic ulcer disease with aspirin, NSAIDs, and potassium. We observed the expected negative associations ($p < 0.05$) of colorectal cancer with aspirin, NSAIDs, and cox-2 inhibitors. The expected protective effect of aspirin use for myocardial infarction was not observed. In all cases, the results obtained from the GPRD™ practices were similar to the results obtained from the non-GPRD™ practices, only being statistically different for the associations of myocardial infarction with diabetes

and aspirin use. CONCLUSIONS: THIN data that are collected outside of the GPRD™ appear as valid as the data collected as part of the GPRD™.

○ Selected Publications By Other Investigators

Gulliford MC, Latinovic R, Charlton J, & Hughes R A C; Increased incidence of carpal tunnel syndrome up to 10 years before diagnosis of diabetes. *Diabetes care 2006; 29(8): 1929-1930.*

A cohort study was implemented using THIN, a database containing medical records from family practices in England and Wales . The study was approved by the South East Research Ethics Committee. Data were analyzed for 114 family practices with 644,495 registered patients aged ≤100 years. Diabetes cases were selected if their diagnosis date was between 1 November 2003 and 31 October 2004 and if they were aged between 30 and 89 years at diagnosis and had never been prescribed insulin or diagnosed with type 1 diabetes. These criteria produced 2,655 cases. Two control groups were randomly selected, matching for age, sex, and practice, from subjects who were never diagnosed with diabetes or prescribed oral hypoglycemic drugs or insulin. Eight cases, for whom two respective control subjects could not be identified, were omitted. Each patient's medical record was searched for first occurrences of CTS, including carpal tunnel release and carpal tunnel injection. We also identified new occurrences of Bell's facial palsy. Other peripheral neuropathies were grouped, including mononeuritis of the upper limb excluding CTS, mononeuritis of the lower limb, mononeuritis multiplex, and peripheral neuropathy (including idiopathic progressive polyneuropathy, polyneuropathy, peripheral neuropathy, other idiopathic peripheral neuropathy, hereditary or idiopathic peripheral neuropathy not otherwise specific, and inflammatory and toxic neuropathies). The entry date to the study was the earlier of the start of the Windows-based medical record or the date of the first prescription following registration (up to a maximum of 10 years before the diabetes diagnosis date).

Mortimer K, Tata LJ , Smith CJP, West J, Harrison TW , Tattersfield A, Hubbard R; (2006). Oral and inhaled corticosteroids and adrenal insufficiency: a case-control study. *Thorax 2006 May; 61(5): 405-8.*

BACKGROUND: Adrenal insufficiency, a well recognised complication of treatment with oral corticosteroids, has been described in association with inhaled corticosteroid use in over 60 case reports. The risk of adrenal insufficiency in people prescribed an oral or inhaled corticosteroid in the general population is not known. A study was undertaken to quantify the association between adrenal insufficiency and oral and inhaled corticosteroid exposure. METHODS: A case-control study was performed using computerised general practice data from The Health Improvement Network. RESULTS: From a cohort of 2.4 million people, 154 cases of adrenal insufficiency and 870 controls were identified. There was a dose related increased risk of adrenal insufficiency in people prescribed an oral corticosteroid with an odds ratio of 2.0 (95% CI 1.6 to 2.5) per course of treatment per year. Adrenal insufficiency was associated with a prescription for an inhaled corticosteroid during the 90 day period before the diagnosis with an odds ratio of 3.4 (95% CI 1.9 to 5.9) and this effect was dose related (p for trend <0.001). After adjusting for oral corticosteroid exposure, this odds ratio was reduced to 1.6 (95% CI 0.8 to 3.2) although the dose relation remained (p for trend 0.036). CONCLUSION: People prescribed an oral or inhaled corticosteroid are at a dose related increased risk of adrenal insufficiency although the absolute risk is small. This analysis suggests that the increased risk in

people prescribed an inhaled corticosteroid is largely due to oral corticosteroid exposure, but inhaled corticosteroids may have an effect when they are taken at higher doses.

Smeeth L, Cook C, Thomas S, Hall AJ, Hubbard R, Vallance P; Risk of deep vein thrombosis and pulmonary embolism after acute infection in a community setting. *Lancet 2006; 367: 1075–79.*

BACKGROUND: Acute infection increases the risk of arterial cardiovascular events, but effects on venous thromboembolic disease are less well established. Our aim was to investigate whether acute infections transiently increase the risk of venous thromboembolism. METHODS: We used the self-controlled case-series method to study the risk of first deep vein thrombosis (DVT) (n=7278) and first pulmonary embolism (PE) (n=3755) after acute respiratory and urinary tract infections. Data were obtained from records from general practices who had registered patients with the UK's Health Improvement Network database between 1987 and 2004. FINDINGS: The risks of DVT and PE were significantly raised, and were highest in the first two weeks, after urinary tract infection. The incidence ratio for DVT was 2.10 (95% CI 1.56-2.82), and that for PE 2.11 (1.38-3.23). The risk gradually fell over the subsequent months, returning to the baseline value after 1 year. The risk of DVT was also higher after respiratory tract infection, but possible diagnostic misclassification precluded a reliable estimate of the risk of PE after respiratory infection. INTERPRETATION: Acute infections are associated with a transient increased risk of venous thromboembolic events in a community setting. Our results confirm that infection should be added to the list of precipitants for venous thromboembolism, and suggest a causal relation.

Van staa TP, Geusens P, Pols HAP, De Laet C, Leufkens HGM; A simple score for estimating the long-term risk of fracture in patients using oral glucocorticoids. *Q J Med 2005; 98:191–198.*

BACKGROUND: Simple tools are needed to identify patients at high risk of fracture. AIM: To develop a simple clinical tool for assessing 5-year risk of fracture. DESIGN: Cohort study. METHODS: The study population consisted of all women aged 50+ included in the THIN Research Database (containing computerized medical records of UK general practices). Using Cox proportional hazards models, a risk score was initially estimated from age, body mass index, and clinical risk factors. The 5-year risk of fracture (survival function) was estimated for each score. RESULTS: The study population included 366 104 women aged > or = 50 years (mean follow-up 5.8 years). Of these, 6453 suffered a hip fracture. Several characteristics independently contributed to the fracture risk score (age, body mass index, fracture and fall history, previous diagnoses and use of medication). The 5-year risks for hip fracture for patients with total scores of 10, 30 and 50 were 0.3% (95%CI 0.3-0.4%), 2.2% (95%CI 2.1-2.2%), and 13.1% (95%CI 12.5-13.7%), respectively. A woman aged 65 years with low BMI and a history of both fracture and falling would have a hip fracture risk score of 37, with a corresponding 5-year risk for a hip fracture of 4.1% (4.0-4.2%). The risk score was validated and tested in another population (from GPRD™), with a good concurrence between predicted and observed risks of fracture. DISCUSSION: This risk score predicts the long-term risk of fracture, and could be used for targeting patients for further investigation, such as bone densitometry.

Arellano FM, Wentworth CE, May C, Verma A, Rivero E, Rothman KJ; Use of cyclo-oxygenase 2 inhibitors (COX-2) and prescription non-steroidal anti-inflammatory drugs (NSAIDS) in UK and USA populations. Implications for COX-2 cardiovascular profile. *PDS 2005;*15(12):861-72.

BACKGROUND: COX-2 and NSAIDS differ in their gastrointestinal (GI) and cardiovascular (CV) toxicity from pharmacological, clinical and epidemiologic point of views. OBJECTIVE: Describe the patterns of use of NSAIDS and COX-2 in The Health Improvement Network (THIN) database in UK and the PharMetrics database in USA. METHODS: We examined the experience of 10 distinct cohorts of new users of diclofenac, naproxen, ibuprofen, piroxicam, other NSAIDS, meloxicam, celecoxib, etoricoxib, rofecoxib and valdecoxib. The study period was 1 January 1995 through 2004 (31 March in UK and 28 February in USA). We collected information on covariates including history of upper GI disease, CV disease, hepatic disease, dosage, concomitant medication, and visits to a rheumatologist. RESULTS: We identified 486 076 unique patient-drug pairs in UK and 1 533 239 in USA. In UK population 78 201 (16%) were COX-2 users and in PharMetrics 324 206 (21%) were COX-2 users. Diclofenac and ibuprofen (NSAIDS), and celecoxib and rofecoxib (COX-2) were the agents prescribed most frequently. The duration of therapy was longer among celecoxib and rofecoxib users than among other users. More COX-2 users than NSAIDS users received concomitant gastroprotective agents (GPA), corticosteroids and anti-platelet therapy, and had a history of thromboembolic events and hypertension. PharMetrics patients were prescribed higher doses of NSAIDS and COX-2. The use of any single agent for more than 90 days was uncommon, but more frequent in PharMetrics. Switching was uncommon and was generally to a NSAID. DISCUSSION: Our results confirm some previous findings from other authors such as the presence of both GI and CV channelling to COX-2 agents but refute others, such as the frequency of drug switching between these agents. The typical use of COX-2 agents in practice is for shorter duration, and at lower doses, than was employed in randomized clinical trials. This difference may help clarify the apparent discrepancy with respect to CV toxicity between the results from clinical trials, which showed a higher CV risk with these drugs, and non-experimental epidemiologic studies, which showed lower or no increase in risk

Hubbard R, Lewis SA, West J, Smith C, Godfrey C, Smeeth L, Farrington P & Britton J; Bupropion and the risk of sudden death: a self-controlled case-series analysis using The Health Improvement Network. *Thorax. 2005 Oct ; 60 (10): 848-50.*

BACKGROUND: Bupropion is an effective smoking cessation therapy but its use in the UK has been limited by concerns that it may increase the risk of sudden death. METHODS: Data for all patients prescribed bupropion within The Health Improvement Network (a computerised general practice database) were extracted and the self-controlled case-series method was used to estimate the relative incidence of death during the first 28 days of treatment. The incidence of seizures, a recognised adverse effect of bupropion, was also investigated during this period. RESULTS: A total of 9329 individuals had been prescribed bupropion (mean age 44 years, 48% male). The total person-time after the first prescription for bupropion was 17,586 years, and during this time 121 people died. Two people died within the first 28 days of treatment, which was less than expected in comparison with the remaining observation period by an incidence ratio of 0.50 (95% confidence interval (CI) 0.12 to 2.05). Twenty eight people were recorded as having a total of 45 seizures (23 before starting bupropion, two in the first 28 days of treatment, and 20 at a later point). The relative incidence of seizures during the first 28 days of treatment was 3.62 (95% CI 0.87 to 15.09), equivalent to one additional seizure per 6219 first time bupropion users. CONCLUSIONS: Bupropion use is probably associated with an increased risk of seizures, but no evidence was found to suggest that the drug is associated with an increased risk of sudden death.

Hubbard R, Lewis S, Godfrey C, Smeeth L, Farrington P; Britton J.Use of nicotine replacement therapy and the risk of acute myocardial infarction, stroke and death. *Tobacco Control 2005, 14(6): 416-421.*

OBJECTIVE: To determine whether nicotine replacement therapy (NRT) is associated with an increased risk of acute myocardial infarction, acute stroke, or death. DESIGN: Self control case series analysis of data from The Health Improvement Network (THIN) to estimate the relative incidence of myocardial infarction and stroke in four 14 day periods before and after the first prescription for NRT. SETTING: THIN is a computerised general practice database. SUBJECTS: Patients contributing data to THIN. INTERVENTIONS: Observational study of NRT. MAIN OUTCOMES: Acute myocardial infarction, acute stroke, and death. RESULTS: 33,247 individuals had been prescribed NRT, of whom 861 had had a myocardial infarction and 506 a stroke. There was a progressive increase in the incidence of first myocardial infarction in the 56 days leading up to the first NRT prescription (overall incidence ratio 5.55, 95% confidence interval (CI) 4.42 to 6.98), but the incidence fell after this time and was not increased in the 56 days after starting NRT (incidence ratio 1.27, 95% CI 0.82 to 1.97). The results were similar for second myocardial infarction and stroke, and for subgroups of people with pre-existing angina and hypertension. There were 960 deaths in our cohort during a mean follow up period of 2.6 years after starting NRT, with no evidence of an increased mortality in the 56 days after the NRT prescription (incidence ratio 0.86, 95% CI 0.60 to 1.23). CONCLUSIONS: The use of NRT is not associated with any increase in the risk of myocardial infarction, stroke, or death.

van Staa TP, Geusens P, Kanis JA, Leufkens HG, Gehlbach S, Cooper C. A simple clinical score for estimating the long-term risk of fracture in post-menopausal women. *QJM.* 2006 Oct;99:673-82.

BACKGROUND: Simple tools are needed to identify patients at high risk of fracture. AIM: To develop a simple clinical tool for assessing 5-year risk of fracture. DESIGN: Cohort study. METHODS: The study population consisted of all women aged 50+ included in the THIN Research Database (containing computerized medical records of UK general practices). Using Cox proportional hazards models, a risk score was initially estimated from age, body mass index, and clinical risk factors. The 5-year risk of fracture (survival function) was estimated for each score. RESULTS: The study population included 366 104 women aged > or = 50 years (mean follow-up 5.8 years). Of these, 6453 suffered a hip fracture. Several characteristics independently contributed to the fracture risk score (age, body mass index, fracture and fall history, previous diagnoses and use of medication). The 5-year risks for hip fracture for patients with total scores of 10, 30 and 50 were 0.3% (95%CI 0.3-0.4%), 2.2% (95%CI 2.1-2.2%), and 13.1% (95%CI 12.5-13.7%), respectively. A woman aged 65 years with low BMI and a history of both fracture and falling would have a hip fracture risk score of 37, with a corresponding 5-year risk for a hip fracture of 4.1% (4.0-4.2%). The risk score was validated and tested in another population (from GPRD™), with a good concurrence between predicted and observed risks of fracture. DISCUSSION: This risk score predicts the long-term risk of fracture, and could be used for targeting patients for further investigation, such as bone densitometry.

Updated 08/25/10