Clinical Epidemiology

Dovepress
open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

The Copenhagen Primary Care Differential Count (CopDiff) database

This article was published in the following Dove Press journal:
Clinical Epidemiology
12 June 2014
Number of times this article has been viewed

Christen Lykkegaard Andersen[1,2]
Volkert Dirk Siersma[1]
Willy Karlslund[1]
Hans Carl Hasselbalch[2]
Peter Felding[3]
Ole Weis Bjerrum[4]
Niels de Fine Olivarius[1]

[1]The Research Unit for General Practice and Section of General Practice, Department of Public Health, University of Copenhagen, [2]Department of Hematology, Roskilde University Hospital, [3]The Elective Laboratory of the Capital Region, [4]Department of Hematology, Copenhagen University Hospital, Copenhagen, Denmark

Background: The differential blood cell count provides valuable information about a person's state of health. Together with a variety of biochemical variables, these analyses describe important physiological and pathophysiological relations. There is a need for research databases to explore associations between these parameters, concurrent comorbidities, and future disease outcomes.

Methods and results: The Copenhagen General Practitioners' Laboratory is the only laboratory serving general practitioners in the Copenhagen area, covering approximately 1.2 million inhabitants. The Copenhagen General Practitioners' Laboratory has registered all analytical results since July 1, 2000. The Copenhagen Primary Care Differential Count database contains all differential blood cell count results (n=1,308,022) from July 1, 2000 to January 25, 2010 requested by general practitioners, along with results from analysis of various other blood components. This data set is merged with detailed data at a person level from The Danish Cancer Registry, The Danish National Patient Register, The Danish Civil Registration System, and The Danish Register of Causes of Death.

Conclusion: This paper reviews methodological issues behind the construction of the Copenhagen Primary Care Differential Count database as well as the distribution of characteristics of the population it covers and the variables that are recorded. Finally, it gives examples of its use as an inspiration to peers for collaboration.

Keywords: differential leukocyte count, research, nationwide health registers

Introduction

One of the most common blood tests in the world, the differential blood cell count (DIFF), provides valuable information on the relative percentage of each type of white blood cell in the peripheral blood. It also provides data on the occurrence of abnormal white blood cell populations like leukemic blast cells, immature myeloid cells, and circulating lymphoma cells. Together with the hemoglobin and platelet count, the DIFF constitutes the complete blood count (CBC), which supplies important information about a person's state of health. Blood sampling may be done in all corners of the health care sector and takes only minutes to perform. Anticoagulants in sample media allow storage of samples for hours, even days if properly cooled.[1] Accordingly, the DIFF and CBC are used for a broad range of medical indications in diagnosis and monitoring of disease activity. Together with a variety of other biochemical parameters, it is also possible to monitor medical therapy as well as to establish prognostic indexes for a plethora of diseases.

However, there is a need for research databases to explore associations between these cellular and biochemical variables, prior and/or concurrent comorbidities, and

Correspondence: Christen Lykkegaard Andersen
The Research Unit for General Practice and Section of General Practice, Department of Public Health, University of Copenhagen, Denmark, Øster Farimagsgade 5, 1014 Copenhagen, Denmark
Tel +45 2612 2840
Fax +45 3532 7131
Email christenla@gmail.com

disease outcomes. Such databases offer unique opportunities to follow long-term outcomes of well-characterized individuals. In 2008, researchers from the Research Unit for General Practice in Copenhagen began constructing the Copenhagen Primary Care Differential Count (CopDiff) database in order to meet this demand.

The CopDiff database extends the purpose of other important research databases such as 1) the Copenhagen General Population Study,[2] since the CopDiff database encompasses both the young (<18 years) and the old (>80 years), and 2) the Clinical Laboratory Information System research database,[3] since the CopDiff supplies data from primary care. By linkage to nationwide health registers, the CopDiff will have the capability to assess the prognostic value of the DIFFs for several clinical outcomes, while adjusting for certain potential confounders. Access to data on some 550,000 individuals (constituting some 10% of the Danish population) over a 10-year period enables the CopDiff database to assess both common and rare disease outcomes. A regular update will allow for the extension of the database into the future with still more outcomes/events.

The purpose of this paper is to review the content of the CopDiff database, describe basic analytical approaches (biochemical and statistical), and lastly, to encourage collaboration with fellow scientists.

Materials and methods

Construction and content

The cellular and biochemical variables of the CopDiff database

In the Copenhagen area, with its approximately 1.2 million inhabitants, there is only one laboratory serving general practitioners (GPs), the Copenhagen General Practitioners' Laboratory (CGPL), known as the Elective Laboratory of the Capital Region since January 1, 2013. The CGPL was founded in 1922 and serves a total of 739 GPs in 567 practices (2010) with a broad range of blood tests, clinical physiological tests, and various cardiac tests. The CGPL has International Organization for Standardization 15189 accreditation and has saved all values on the analyses it performs since July 2000. The CGPL offers two routine groups of hematology analyses for the GPs:

1. "HEM": hemoglobin, mean red cell volume, red cell distribution width, total white blood cell count, and platelet count.
2. "CBC": the HEM group plus differential counts of white blood cells (neutrophils, lymphocytes, monocytes, eosinophils, and basophils).

The individual components of the groups cannot be requested alone. To obtain, for example, hemoglobin, the GP has to request either "HEM" or "CBC".

The CBC requests from the period July 1, 2000 until January 25, 2010 were included in the first step of building the CopDiff database (Table 1). The stand-alone "HEM" requests were excluded. All other analyses requested by the GP in addition to the CBC, if on the same requisition, were also included in the database (Table 2). Hence, common for all individuals was the existence of a CBC estimation while the remaining analyses were only included for a particular patient if the GP had ordered these analyses on the same requisition on which the CBC was ordered. Requests for CBCs from non-GPs (ie, specialized consultants with their own practices) were excluded in order to obtain a pure primary care resource (Figure 1). Of note, these requisitions have not been deleted from the CopDiff servers and may be analyzed if it becomes relevant to include them in an analysis.

The CopDiff database eventually included 1,308,022 requisitions on 555,039 unique individuals to be further merged with data from nationwide registers described below. All requisitions with numeric and alphanumeric (but valid) results were also categorized according to reference limits at the time of the analysis as either normal, below, or above reference range in separate variables (Table 2).

Analytical methods of the CopDiff database

All CBC samples were analyzed on Siemens (Bayer/Technicon, Munich, Germany) hematology systems. CGPL used three similar types of these instruments in the period 2000–2010, which in chronological order were Technicon® H3 RTX (used between 2000 and 2002), ADVIA® 120 (used between 2002 and 2010), and ADVIA® 2120i (used together with ADVIA® 120 from 2009 to 2010). The basic chemical and physical methods are identical among these systems. In general, samples were treated with certain chemicals

Table 1 Characteristics of the CopDiff database population

Sex, n (%)	
Male	232,251 (41.8)
Female	322,788 (58.2)
Age at first requisition, years	46.9±21.5
Requisitions, total	1,308,022
Requisitions per patient	2.36±2.81
Deaths before January 25, 2010	61,416 (11.1)
Emigrated/disappeared/inactive before January 25, 2010	10,669 (1.9)
Years from first requisitions until January 25, 2010 or death/emigration/inactivation	4.98±2.87

Note: Values are numbers (%) or means (SD).
Abbreviations: CopDiff, Copenhagen Primary Care Differential Count; SD, standard deviation.

Table 2 All requisitions in the CopDiff database sorted by prevalence in regard to requisitions with numeric results

Name (IUPAC[1] code), unit	Requisitions with numeric results (% of all requisitions)[2]	Minimum	1st quartile	Median	3rd quartile	Maximum	Normal (%)[3]	Below reference (%)[3]	Above reference (%)[3]	Missing/ failed (%)[3]	Cancelled (%)[3]
Erythrocytes (NPU01944), (MCV, fL)	1,307,863 (100)	45.0	88.0	91.0	95.0	148.0	1,192,306 (91.2)	33,520 (2.6)	82,037 (6.2)	155 (0.0)	4 (0.0)
Hemoglobin (NPU02319), mmol/L	1,307,860 (100)	1.3	7.9	8.5	9.1	13.9	1,185,879 (90.6)	115,898 (8.9)	6,091 (0.5)	118 (0.0)	36 (0,0)
Erythrocyte volumes; relative distribution width (NPU18162), %	1,307,857 (100)	4.3	12.8	13.3	13.9	51.1	1,199,270 (91.7)	53 (0.0)	108,534 (8.3)	159 (0.0)	6 (0.0)
Leukocytes (NPU02593), 10^9/L	1,307,830 (100)	0.0[4]	5.8	7.1	8.8	695.0	1,019,971 (78.0)	8,170 (0.6)	279,689 (21.4)	185 (0.0)	7 (0.0)
Thrombocytes (NPU03568), 10^9/L	1,305,783 (100)	1.0[4]	228.0	273.0	326.0	3,369.0	1,161,370 (88.8)	34,551 (2.6)	111,861 (8.6)	235 (0.0)	5 (0.0)
Lymphocytes (NPU02636), 10^9/L	1,292,663 (98.8)	0.0[4]	1.6	1.98	2.5	650.0	1,261,655 (96.5)	17,286 (1.3)	13,736 (1.0)	15,345 (1.2)	0 (0.0)
Monocytes (NPU02840), 10^9/L	1,292,302 (98.8)	0.0[4]	0.3	0.4	0.6	67.0	1,278,125 (97.7)	17 (0.0)	14,528 (1.1)	15,350 (1.2)	2 (0.0)
Neutrophilocytes (NPU02902), 10^9/L	1,283,902 (98.2)	0.0[4]	3.2	4.2	5.5	95.9	1,141,770 (87.3)	28,464 (2.2)	118,925 (9.0)	10,131 (0.8)	8,732 (0.7)
Eosinophilocytes (NPU01933), 10^9/L	1,279,654 (96.7)	0.0[4]	0.1	0.2	0.3	29.8	1,222,197 (92.4)	694 (0.0)	85,125 (6.4)	15,347 (1.2)	0 (0.0)
Basophilocytes (NPU01349), 10^9/L	1,212,203 (92.7)	0.00[4]	0.03	0.04	0.06	16.30	1,289,876 (98.6)	0 (0.0)	1,983 (0.2)	16,162 (1,2)	1 (0.0)
Creatininium (NPU01807), µmol/L	971,928 (100)	16.0	78.0	88.0	99.0	1,645.0	904,066 (93.0)	1,146 (0.0)	66,722 (6.9)	1,675 (0.1)	290 (0.0)
Alanine transaminase (NPU19651), U/L	874,961 (99.8)	3.0	17.0	23.0	33.0	10,580.0	779,297 (88.9)	10,589 (1.2)	85,177 (9.7)	1,003 (0.1)	857 (0.1)
Sodium ion (NPU03429), mmol/L	756,418 (99.7)	103.0	139.0	141.0	142.0	184.0	704,544 (92.8)	42,254 (5.6)	9,620 (1.3)	2,539 (0.3)	181 (0.0)
Potassium ion[A] (NPU03230), mmol/L	742,539 (97.8)	1.6	4.1	4.3	4.5	9.8	702,319 (92.5)	11,683 (1.5)	28,538 (3.8)	3,818 (0.5)	10,405 (1.4)
Thyrotropin (NPU03577), $\times 10E^{-3}$ IU/L	717,803 (98.5)	0.02	0.95	1.4	2.1	150.0	674,573 (92.6)	21,206 (2.9)	30,738 (4.2)	1,668 (0.3)	365 (0.0)
Alkaline phosphatase† (NPU19655), U/L	716,051 (99.8)	2.0	65.0	89.0	158.0	15,770.0	629,331 (87.7)	8,409 (1.2)	78,318 (10.9)	1,299 (0.2)	132 (0.0)
Cholesterol + ester (NPU01566), mmol/L	522,401 (99.8)	0.8	4.6	5.4	6.2	43.0	445,224 (85.1)	63,063 (12.0)	14,116 (2.7)	1,223 (0.2)	26 (0.0)
Cholesterol + ester (NPU01567) in HDL, mmol/L	433,568 (99.7)	0.2	1.1	1.4	1.7	7.0	394,387 (90.7)	4,558 (1.0)	34,643 (8.1)	1,032 (0.2)	71 (0.0)

(*Continued*)

Table 2 (*Continued*)

Name (IUPAC[1] code), unit	Requisitions with numeric results (% of all requisitions)[2]	Minimum	1st quartile	Median	3rd quartile	Maximum	Normal (%)[3]	Below reference (%)[3]	Above reference (%)[3]	Missing/ failed (%)[3]	Cancelled (%)[3]
C-reactive protein[(B)] (NPU19748), mg/L	376,815 (46.1)	0.0[4]	6.0	10.0	21.0	400.0[4]	634,599 (77.6)	0 (0.0)	181,524 (22.2)	1,845 (0.2)	289 (0.0)
Triglyceride (NPU03620), mmol/L	362,525 (99.8)	0.2	0.9	1.3	1.9	118.0	296,155 (81.6)	5,985 (1.6)	60,388 (16.6)	795 (0.2)	93 (0.0)
Cholesterol + ester[(C)] (NPU10171) in LDL, mmol/L	351,589 (97.7)	–0.60[x]	2.6	3.2	3.9	16.4	299,437 (83.1)	41,697 (11.6)	10,455 (2.9)	938 (0.3)	0 (0.0)
Hemoglobin A_{1c} (Fe) (NPU03835), %	223,785 (99.3)	0.0[4]	5.5	5.8	6.5	18.9	164,868 (73.2)	0 (0.0)	59,051 (26.2)	997 (0.4)	342 (0.2)
Albumin (NPU19673), g/L	201,248 (99.7)	9.9	40.7	42.7	44.5	71.0	183,762 (91.0)	15,930 (7.9)	1,556 (0.8)	565 (0.3)	43 (0.0)
Sedimentation reaction (NPU03404), mm	190,188 (98.9)	0.0[4]	5.0	10.0	19.0	150.0	142,876 (74.1)	1,401 (0.7)	47,396 (24.6)	706 (0.4)	339 (0.2)
Iron[(D)] (NPU02508), µmol/L	189,307 (94.5)	2.0	10.0	15.0	19.0	89.0	146,428 (73.1)	41,685 (20.8)	2,216 (1.1)	434 (0.2)	368 (0.2)
Cobalamin (NPU01700), pmol/L	176,111 (95.0)	35.0	240.0	305.0	395.0	1,500.0	175,103 (94.4)	4,354 (2.3)	127 (0.1)	647 (0.3)	5,320 (2.9)
Glucose[(E)] (KPL00290), mmol/L	164,577 (94.0)	0.6	4.9	5.4	6.0	54.0	130,526 (74.6)	3,004 (1.7)	31,052 (17.7)	1,011 (0.6)	1,233 (0.7)
Ferritin (NPU03899), pmol/L	154,790 (99.5)	1.0	29.0	67.0	147.0	14,900.0	127,321 (81.9)	13,274 (8.5)	14,410 (9.3)	334 (0.2)	153 (0.1)
Glucose[(F)] (KPL00291), mmol/L	148,686 (88.0)	0.6	4.8	5.2	5.9	48.2	129,677 (76.7)	9,236 (5.5)	9,792 (5.8)	1,917 (1.1)	2,598 (1.5)
Glucose[(G)] (NPU02192), mmol/L	129,212 (98.0)	1.0	4.9	5.5	6.4	72.8	108,530 (82.3)	5,752 (4.4)	14,932 (11.2)	562 (0.4)	1,690 (1.3)
Glucose[(G)] (NPU02195), mmol/L	124,458 (98.3)	0.9	5.1	5.6	6.4	46.8	87,461 (69.1)	1,386 (1.1)	35,611 (28.1)	376 (0.3)	1,328 (1.0)
Thyroxine, free (NPU03579), pmol/L	93,632 (99.6)	3.0	13.3	15.1	17.4	144.0[4]	84,580 (90.1)	3,237 (3.4)	5,936 (6.3)	229 (0.2)	16 (0.0)
Transferrin (NPU03607), µmol/L	92,405 (99.6)	7.0	28.0	32.0	36.0	76.0	78,297 (84.4)	5,535 (6.0)	8,581 (9.3)	319 (0.3)	25 (0.0)
Transferrin (Fe-binding sites; P)-Iron; subst fr = ? or transferrin saturation[(H)], Fe-binding sites (NPU04191)	79,460 (94.0)	0.01	0.15	0.22	0.30	1.46	62,327 (73.4)	15,461 (18.2)	1,672 (2.0)	20 (0.0)	0 (0.0)
Prostate specific antigen (NPU08669), µg/L	70,878 (86.4)	0.1	0.7	1.4	3.1	970.0	68,233 (83.2)	0 (0.0)	13,501 (16.5)	284 (0.3)	22 (0.0)
Triiodothyronine (NPU03624), nmol/L	68,907 (99.8)	0.0[4]	1.6	1.8	2.2	23.2	62,119 (89.9)	791 (1.1)	5,998 (8.9)	148 (0.1)	9 (0.0)
Thyroxine (NPU03578), nmol/L	60,367 (99.5)	0.0[4]	90.0	105.0	123.0	460.0	55,709 (91.8)	1,493 (2.5)	3,318 (5.5)	128 (0.2)	5 (0.0)

Epstein–Barr virus capsid antibody, Immunoglobulin M (NPU12738), ELISA	45,583 (97.2)	0.0	0.0	0.0	0.0	2.0	38,440 (82.0)	0 (0.0)	8,457 (18.0)	6 (0.0)	0 (0.0)
Immunoglobulin A (NPU19795), g/L	41,765 (99.4)	0.0[4]	1.6	2.2	2.9	81.2	38,524 (91.8)	642 (1.5)	2,685 (6.4)	139 (0.3)	14 (0.0)
Rheumatoid factor antibody[(I)] (NPU18350), 10^3 IU/L	40,466 (45.6)	3.0	7.0	11.0	14.0	250.0	79,336 (89.4)	0 (0.0)	9,136 (10.3)	252 (0.3)	20 (0.0)
Reticulocytes (NPU08694), 10^9/L	40,254 (99.4)	0.1[4]	46.3	58.5	73.7	649.2	38,277 (94.6)	703 (1.7)	1,285 (3.2)	201 (0.5)	13 (0.0)
Immunoglobulin M (NPU19825), g/L	39,488 (99.5)	0.0[4]	0.7	0.96	1.3	60.3	35,213 (88.8)	2,123 (5.4)	2,194 (5.5)	136 (0.3)	10 (0.0)
Immunoglobulin G (NPU19814), g/L	39,399 (99.6)	1.0	9.0	10.6	12.5	118.0	35,291 (89.2)	756 (2.0)	3,357 (8.4)	140 (0.4)	0 (0.0)
Glucose[(J)] (DNK35842), mmol/L	35,466 (94.7)	2.7	5.3	5.7	6.4	37.3	25,475 (68.0)	105 (0.3)	9,886 (26.4)	139 (0.4)	1,509 (4.0)
Reticulocyte-hemoglobin (NPU17007), fmol	34,429 (99.4)	0.95	1.8	1.96	2.1	3.2	27,585 (79.8)	4,726 (13.6)	2,118 (6.1)	179 (0.5)	24 (0.0)
Parathyrin (NPU03028), pmol/L	32,917 (99.1)	0.3	3.2	4.5	6.5	165.0	26,786 (80.6)	706 (2.1)	5,478 (16.5)	95 (0.3)	156 (0.5)
Neutrophilocytes, segmented[(K)] (NPU03982), 10^9/L	28,093 (54.2)	0.0	2.2	3.4	5.4	85.3	19,956 (38.6)	3,900 (7.5)	4,237 (8.2)	6 (0.0)	0 (0.0)
Neutrophilocytes, band[(L)] (NPU03980), 10^9/L	28,085 (54.2)	0.0	0.0	0.0	0.1	37.3	25,596 (49.4)	0 (0.0)	2,489 (4.8)	6 (0.0)	0 (0.0)
Lactate dehydrogenase[†] (NPU19658), U/L	26,334 (100)	72.0	270.0	324.0	378.0	4,527.0	23,012 (82.6)	69 (0.2)	3,253 (11.7)	1,406 (5.0)	142 (0.5)
Urine albumin/creatininium[(M)] (NPU19661), mg/g	18,663 (46.7)	0.0[4]	15.0	27.0	71.0	85,400	29,151 (72.9)	0 (0.0)	8,637 (21.6)	518 (1.3)	1,690 (4.2)
Mononucleosis reaction (NPU03946), 0/1 (negative/positive)	17,554 (99.7)	0.0	0.0	0.0	0.0	1.0	14,956 (85.0)	0 (0.0)	2,598 (14.7)	6 (0.0)	54 (0.3)
Orosomucoid (NPU19873), g/L	13,683 (99.7)	0.0[4]	0.7	0.9	1.1	4.2	11,014 (80.2)	147 (1.1)	2,522 (18.4)	42 (0.3)	6 (0.0)
Brain natriuretic peptide, from November 2, 2006 (NPU17181), pmol/L	13,598 (96.8)	0.6	4.1	9.0	22.2	1,311.0	11,194 (79.7)	0 (0.0)	2,669 (19.0)	103 (0.7)	84 (0.6)
Folate (NPU02070), nmol/L	13,245 (92.5)	0.8	10.5	15.8	24.6	54.3	13,453 (94.0)	798 (5.6)	0 (0.0)	33 (0.2)	34 (0.2)
Nuclear antibody, Immunoglobulin G (NPU14127), pdu	9,408 (98.3)	0.1	0.2	0.3	0.4	30.0	9,072 (94.8)	17 (0.2)	390 (4.1)	87 (0.9)	8 (0.0)
DNA, double stranded antibody Immunoglobulin G (NPU16393), 10E3 IU/L	7,179 (75.2)	0.6	0.9	1.4	2.7	167.0	9,036 (94.7)	0 (0.0)	415 (4.3)	75 (0.8)	21 (0.2)

(*Continued*)

Table 2 (*Continued*)

Name (IUPAC[1] code), unit	Requisitions with numeric results (% of all requisitions)[2]	Minimum	1st quartile	Median	3rd quartile	Maximum	Normal (%)[3]	Below reference (%)[3]	Above reference (%)[3]	Missing/ failed (%)[3]	Cancelled (%)[3]
Carcinoembryonic antigen (NPU19719), µg/L	3,198 (70.2)	0.5[4]	1.6	2.3	3.8	988.0	4,208 (92.3)	0 (0.0)	333 (7.3)	15 (0.3)	0 (0.0)
Probrain natriuretic peptide, 1–76 (NPU26811), pmol/L	2,887 (97.0)	1.0	10.0	26.0	86.0	4030	1,624 (54.5)	0 (0.0)	1,268 (42.6)	68 (2.3)	17 (0.6)
Blast cells, unspecified[(N)] (NPU03972), 10^9/L	2,598 (5.0)	0.0	0.0	0.0	0.0	496.0	2,025 (3.9)	0 (0.0)	573 (1.1)	6 (0.0)	0 (0.0)
Promyelocytes[(N)] (NPU03974), 10^9/L	2,567 (5.0)	0.00	0.00	0.00	0.02	38.20	1,914 (3.7)	0 (0.0)	653 (1.3)	6 (0.0)	0 (0.0)
Metamyelocytes[(N)] (NPU03978), 10^9/L	2,562 (5.0)	0.0	0.0	0.0	0.2	16.3	1,409 (2.7)	0 (0.0)	1,153 (2.2)	6 (0.0)	0 (0.0)
Troponin I, cardiac muscle[(O)] (NPU19923), µg/L	2,109 (27.3)	0.01	0.01	0.02	0.03	30.3	6,854 (88.7)	0 (0.0)	333 (4.3)	73 (0.9)	470 (6.1)
Haptoglobin (NPU19788), g/L	441 (80.6)	0.2	0.9	1.4	1.9	5.9	316 (57.8)	52 (9.5)	101 (18.5)	40 (7.3)	38 (6.9)
Plasmocytes[(P)] (NPU04708), 10^9/L	122 (0.2)	0.00	0.00	0.00	0.23	2.02	75 (0.2)	0 (0.0)	47 (0.0)	6 (0.0)	0 (0.0)
Inhalation antigen-Ab, Immunoglobulin E[(Q)]	0 (0.0)	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a

Notes: [1]The International Union of Pure and Applied Chemistry (IUPAC) codes provide the terminology for properties and units in the clinical laboratory sciences;[9] [2]represents requisitions with numeric results in less than 90% of total requisitions; indicators of distribution such as minimum, maximum, and median may be biased; [3]represents % of all requisitions; [4]due to extraordinary manual procedures in the laboratory, this result has escaped an algorithmic transformation to below report limit or to higher than report limit; [†]the property defined by this NPU code was introduced March 11, 2004; before that, alkaline phosphatase and lactate dehydrogenase were measured with methods giving higher results; these higher results are included in the CopDiff database and will be taken into account when the results are used in investigations; [(A)]0.3% of requisitions failed due to hemolysis; [(B)]53.9% of requisitions are alphanumeric (below report limit); [(C)]triglycerides >4.5 mmol/L =2.1%; [(D)]4.6% of requisitions failed due to hemolysis; [(E)]result may be found in another available glucose requisition =4.7%; [(F)]result may be found in another available glucose requisition =9,4%; [(G)]result may be found in another available glucose requisition =0.4%; [(H)]6% of requisitions not possible to calculate; [(I)]54.4% of requisitions being alphanumeric (below report limit); [(J)]result may be found in another available glucose requisition =0.9%; [(K)]45.7% of requisitions are alphanumeric (no immature forms); [(L)]45.8% of requisitions are alphanumeric (no immature forms); [(M)]47.8% of requisitions are alphanumeric (below report limit); [(N)]95.0% of requisitions are alphanumeric (no immature forms); [(O)]20.0% of requisitions being alphanumeric (below report limit); [(P)]99.8% of requisitions are alphanumeric (no immature forms); [(Q)]negative =60.3%, positive =39.4%; *calculated value: total-cholesterol – HDL – 0.45 × triglycerides; a negative result indicates that the formula is not applicable in abnormally low values of cholesterol, HDL, or triglycerides.

Abbreviations: ELISA, enzyme-linked immunosorbent assay; HDL, high density lipoprotein; MCV, mean corpuscular volume; n/a, not applicable; NPU, Nomenclature, Properties, and Units; LDL, low density lipoprotein; CopDiff, Copenhagen Primary Care Differential Count; subst fr, substance fraction; pdu, procedure defined unit.

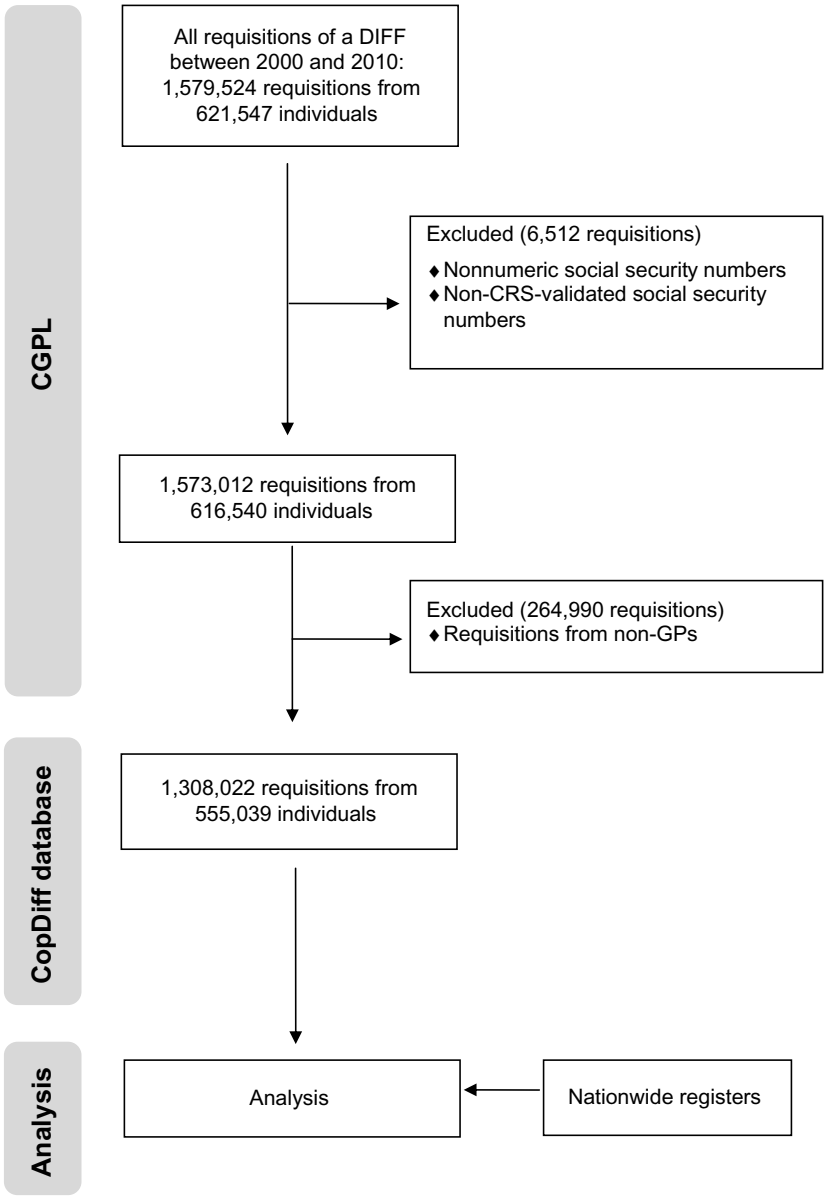

Figure 1 Flowchart.
Abbreviations: CGPL, Copenhagen General Practitioners' Laboratory; CopDiff, Copenhagen Primary Care Differential Count; CRS, Danish Civil Registration System; DIFF, differential cell count; GP, general practitioner.

inside the instruments and the absorbance of hemolysate and light scatter of individual cells were measured. Samples were subjected to microscopic (manual) differential cell counting of leukocyte types if flagged for this during the initial automated differential counting. For this, we used polychrome methylene blue and eosin stains. Method principles valid for all three systems, based on the ADVIA® 120 system, are accessible in the Supplementary materials (Appendix 1 and Table S1). When switching from H3 RTX to ADVIA® 120, there was a relative drop of 5% in "red cell distribution width" analyses. Reference intervals were updated accordingly. No other changes in hematological analyses in the CopDiff period (2000–2010) were performed. The properties defined by the Nomenclature, Properties, and Units codes for alkaline phosphatase and lactate dehydrogenase, as noted in Table 2, were introduced March 11, 2004. Before that, alkaline phosphatase and lactate dehydrogenase were measured with methods that gave higher results. These higher results are also included in the CopDiff database.

Danish nationwide registers on health and social status

Denmark has a long tradition of collecting miscellaneous information on disease incidence, social relations, and

other data describing its population. Permanent residents in Denmark are provided with a personal identification number, which functions as a cornerstone in efficient linkage between all registers containing information at an individual level. Existing registers on health issues encompass, among other information, data on the use of primary and secondary health care as well as diagnoses from contacts with hospitals, including psychiatric hospitals, benign and malignant conditions, and the prescription of drugs in primary health care. Furthermore, available registers on social issues contain data on education, living conditions, labor, earnings, income, etc.[4] Researchers employed at authorized research institutions in Denmark can obtain access to individual level data which enables the Research Unit for General Practice in Copenhagen to link paraclinical data from the CGPL to nationwide registers. A comprehensive list of existing registers, including detailed information on structure, access, legislation, and archiving of Danish registers on health and social issues, has been reviewed recently.[4] Also, a thorough description of the most important Danish population-based registers for public health and health-related welfare research has been published.[5]

In April 2011 (and again in November 2013), the CopDiff database linked all 555,039 individuals to 1) the Danish Civil Registration System (CRS); 2) the Danish Cancer Registry, containing data on all malignancies in Denmark since 1942 and to which reporting is mandatory;[6] 3) the Danish National Patient Register including information on all contacts with hospitals in Denmark, inclusive of discharge diagnoses, outpatient clinic contacts, and surgical procedures performed;[7] and 4) the Register of Causes of Death, containing information on causes of death based upon death certificates.[8]

Results

Utility

The general type of research question that can be answered by the CopDiff database is whether certain levels of selected blood components are associated with an increased risk of certain future disease outcomes. Given that the CopDiff database was constructed on the basis of existing DIFFs, any researcher with a hypothesis not directly involving leukocytes as main variables of interest should bear in mind the risk of inappropriately excluding potential relevant individuals if such cases have not been referred to DIFF sampling by their GPs. Notably, DIFF sampling performed or requested in secondary care is not included, and the CopDiff database therefore does not contain hospitalized individuals. Furthermore, due to inclusion/selection of individuals who have been referred to DIFF sampling, it may be assumed that the CopDiff cohort has more morbidity than the (nonhospitalized) background population. Statistical methods implemented to analyze these data have to take this selection bias into account. Another challenge is the assessment of outcomes in the presence of competing risks. Particular leukocyte configurations may increase mortality, reducing the time for certain diseases of interest to develop, and thereby artificially reduce the risk for such diseases. In the Supplementary materials (Appendix 2), we give a portfolio of statistical analysis designs and their advantages and disadvantages and illustrate their use with data from an already published study.

Discussion

By constructing the CopDiff database, we believe a novel opportunity has been created to explore associations between DIFFs from the peripheral blood and biochemical parameters, concurrent comorbidities, and disease outcomes. An important limitation in the construction of the database is the way the individuals were selected in general practice, with a wide variety of unknown clinical problems, and individuals without existing DIFFs were not included. Nevertheless, by encompassing the young (<18 years) and the old (>80 years), the CopDiff database allows for the assessment of associations through a lifetime for a large primary care population. The access to all DIFFs from all GPs in the Copenhagen area over a 10-year period offers unique insight into the entire Copenhagen area, covering approximately 1.2 million inhabitants. Since the CopDiff population was sampled continuously without any restrictions as to why the DIFF was requested by the GP, the risk of selection bias is diminished among these individuals. In time, the merging of the CopDiff database with other population-based registers such as the Danish Drug Prescription Register, the Danish Heart Register, the National Diabetes Register, and the Danish Psychiatric Central Research Register will allow for exploration of new areas of research. The possible combined assessment of individuals from the general population (Copenhagen General Population Study), primary care (CopDiff), and secondary care (Clinical Laboratory Information System) will provide the basis for unique insight into patient journeys.

Conclusion

This paper has given insight into the fundamentals of the CopDiff database, described its content, and by giving

examples of statistical analytical approaches, hopefully inspired researchers to develop possible future uses. We hereby encourage our peers to contact us in order to collaborate on new projects or to test hypotheses in the CopDiff cohort.

Further information on the CopDiff database, steering committee, bylaws, and ways to collaborate can be found by visiting the CopDiff homepage (http://almenpraksis.ku.dk/english/research/copdiff/).

Author contributions

CLA, WK, VDS, PF, and NdFO designed the CopDiff database, collected, analyzed, and interpreted data and drafted the manuscript. VDS performed the statistical analyses. HCH and OWB analyzed and interpreted data. All authors critically revised the manuscript for important intellectual content and approved the final version to be submitted. All authors agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work were appropriately investigated and resolved.

Acknowledgments

CLA wishes to thank the Danish Cancer Society, which has granted a 3-year scholarship (2010–2013). The authors would also like to express their gratitude to the Eva and Henry Frænkels' Memorial Foundation and the Axel Muusfeldts Memorial Foundation for financial support.

Disclosure

The study has received no financial support or other benefits from commercial sources, and none of the authors have any financial interests which could create potential conflicts of interest. The authors report no conflicts of interest in this work.

References

1. Cornet E, Behier C, Troussard X. Guidance for storing blood samples in laboratories performing complete blood count with differential. *Int J Lab Hematol*. 2012;34:655–660.
2. Nordestgaard BG, Tybjaerg-Hansen A, Fogh-Andersen N, et al. The Copenhagen General Population Study [webpage on the Internet]. Available from: http://www.cgps.dk/SteeringCommittee.php. Accessed May 18, 2014.
3. Grann AF, Erichsen R, Nielsen AG, Frøslev T, Thomsen RW. Existing data sources for clinical epidemiology: The clinical laboratory information system (LABKA) research database at Aarhus University, Denmark. *Clin Epidemiol*. 2011;3:133–138.
4. Thygesen LC, Daasnes C, Thaulow I, Bronnum-Hansen H. Introduction to Danish (nationwide) registers on health and social issues: structure, access, legislation, and archiving. *Scand J Public Health*. 2011; 39(Suppl 7):12–16.
5. Thygesen LC, Ersbøll AK. Danish population-based registers for public health and health-related welfare research – A description of Danish registers and results from their application in research. *Scand J Public Health*. 2011;39(Suppl 7):8–10.
6. Gjerstorff ML. The Danish Cancer Registry. *Scand J Public Health*. 2011;39(Suppl 7):42–45.
7. Lynge E, Sandegaard JL, Rebolj M. The Danish National Patient Register. *Scand J Public Health*. 2011;39(Suppl 7):30–33.
8. Juel K, Helweg-Larsen K. The Danish registers of causes of death. *Dan Med Bull*. 1999;46(4):354–357.
9. International Union of Pure and Applied Chemistry [homepage on the Internet]. Available from: http://www.iupac.org/. Accessed May 18, 2014.

Supplementary materials

Appendix 1

CBC, reticulocyte, and white blood cell differential analyses using the ADVIA® 120 Hematology System

The ADVIA® 120 was used from 2002 to 2010, and the method principles for this instrument are valid for all three instrument types (Technicon® H3 RTX, ADVIA® 120 and ADVIA® 2120i). We used the cyanide hemoglobin method throughout the period (also on ADVIA® 2120i). Basically, the samples were treated with chemicals in the instruments, and the absorbance of hemolysate and light scatter of individual cells were measured.[1]

Principle of the test

The ADVIA® 120 Hematology System is a fully automated diagnostic instrument that uses cytochemical reactions to differentiate and count white blood cells, red blood cells, platelets, and reticulocytes. There are two main components of the system: the analyzer and the personal computer. In the analyzer, blood samples are aspirated and divided into aliquots for the different types of tests. Reagents and segmented samples are delivered to reaction chambers where they are mixed, and a cytochemical reaction takes place. Once the reactions are complete, the sample and reagent mixtures from the so-called "peroxidase", "red blood cell", "basophil", and "reticulocyte" reaction chambers are sent to the flowcells for analysis. The hemoglobin measurement is read in the hemoglobin reaction chamber that serves as an optical cuvette. After analysis, the sample and reagent mixture are evacuated into the waste container and the appropriate pathways and reaction chambers are rinsed. Test results are sent to the computer to be reviewed and edited.

The ADVIA® 120 Hematology System can run five selectivities: CBC, CBC/DIFF, reticulocytes, CBC/reticulocytes, and CBC/DIFF/reticulocytes. The system has a throughput of 120 samples per hour when running CBC or CBC/DIFF and a throughput of 74 samples per hour when running the other selectivities. Up to 150 sample tubes can be loaded onto the barcoded racks of the autosampler. Single or STAT (short turn around time) samples can be tested on the manual samplers (Table S1).

Appendix 2

Design considerations and case study: eosinophilia in routine blood samples and the subsequent risk of hematological malignancies[2]

The nature of the data in the CopDiff database – the way the data are obtained, the dynamics of the capture population, the sheer amount of data – requires careful considerations regarding the methods used to analyze relevant hypotheses. Two basic analytical approaches that can be considered are the case-control design and the cohort design.

Case-control designs

In a classic case-control design, we take the outcome of interest as our point of departure, and cases are the individuals who experience this outcome. Exposure for the cases is then determined by the latest measurement in the CopDiff database within a fixed period before the (first) occurrence of the outcome. For each case, controls have to be chosen from all individuals who do not have the outcome at the time the case's outcome occurs. Choosing all controls for each case will cause controls to feature in the data multiple times. We need to control for this feature in the analysis or in some clever linkage of controls to cases. Choosing a limited number of controls, possibly matched on some characteristics of the case, will reduce the data and reduce, but not solve, the multiplicity problem. Moreover, controls will feature in the data only when a measurement in the CopDiff database is within the fixed period before the corresponding case's occurrence of the outcome. In conclusion, we find this approach too cumbersome and not suited to answer apparent research questions in the CopDiff database.

Cohort designs

Two other approaches start from the exposure and construct cohort data. We opt for choosing randomly one single measurement in the CopDiff database for each individual in order not to have to control for people that enter the cohort multiple times at different points in time. From the time of the exposure, we then look forward in the Danish national registers for the first occurrence of the outcome. Individuals for whom the outcome has already occurred at the time of the exposure measurement are excluded from the analysis. This information can be used in two ways:

- The first approach uses logistic regression to model the probability of experiencing the outcome in a specified time period after the exposure. The main advantages of this methodology are 1) the outcome is well-understood and answers a clinically relevant question: "Will the risk of experiencing the outcome in the coming x-year period be higher if measurement y, taken now at the laboratory, is abnormal?" and 2) the effect estimate of the exposure is an odds ratio (OR), which is approximately the same as a relative risk and also invariant to the prevalence of the outcome. As mentioned previously, the CopDiff sample is

Table S1 Chemical principles

Test name	Chemical principle
White blood cell count	The whole blood sample is mixed with ADVIA® 120 BASO reagent that contains acid and surfactant. The red cells are hemolyzed, and the white blood cells are then analyzed using two angle laser light scatter signals.
Red blood cell/platelet count	Both red blood cells and platelets are analyzed by a single optical cytometer after appropriate dilution of the blood sample with ADVIA® 120 RBC/PLT reagent. The red blood cells are isovolumetrically sphered and lightly fixed with glutaraldehyde to preserve the spherical shape. Red cells and platelets are counted from the signals from a common detector with two different gain settings. On the ADVIA® 120 Hematology System, the platelet signals are amplified considerably more than the red blood cell signals. Coincidence correction is made to each of the counts so that accurate counts are made over a wide range of each cell type.
Red blood cell/platelet size	The method of sizing red cells and platelets uses the simultaneous measurement of laser light scattered at two different angular intervals, which eliminates the adverse effect of variation in cellular hemoglobin concentration on the determination of cell volume.
Hemoglobin concentration	The hemoglobin method is a modification of the manual cyanmethemoglobin method developed by the International Committee for Standardization in Hematology. The sample and ADVIA® 120 HGB reagent are mixed in the hemoglobin reaction chamber (colorimeter). The hemoglobin chemical reactions consist of two steps: the red blood cells are lysed to release hemoglobin and the heme iron in the hemoglobin is oxidized from the ferrous to the ferric state. It is then combined with cyanide in the ADVIA® 120 HGB reagent to form the reaction product.
Reticulocyte cell count	This method uses a nucleic acid dye (oxazine 750) to stain cellular RNA. Two microliters of an EDTA anticoagulated whole-blood sample are mixed online with the ADVIA® 120 autoRETIC reagent. The ADVIA® 120 autoRETIC reagent isovolumetrically spheres the erythroid cells and stains cellular RNA. Low-angle laser light scatter, high-angle laser light scatter, and absorption characteristics of all cells are counted and measured. The absorption data are used to classify each cell as a reticulocyte or mature red blood cell based on its RNA content.
Reticulocyte size	The method of sizing reticulocytes uses the simultaneous measurement of laser light scattered at two different angular intervals, which eliminates the adverse effect of variation in cellular hemoglobin concentration on the determination of the mean reticulocyte volume parameter.
CHr	The CHr is the mean of cellular hemoglobin content (CH) histogram for the reticulocyte population.
Peroxidase method	The peroxidase cytochemical reaction consists of two steps. In the first step, EDTA anticoagulated whole-blood sample is diluted with ADVIA® 120 PEROX 1 reagent. Surfactants and thermal stress cause lysis of the red blood cells. Formaldehyde in ADVIA® 120 PEROX 1 reagent fixes the white blood cells. During the second step, ADVIA® 120 PEROX 2 reagent and ADVIA® 120 PEROX 3 reagent are added to the peroxidase reaction chamber. The 4-chloro-1-naphthol in ADVIA® 120 PEROX 2 reagent and the hydrogen peroxide in ADVIA® 120 PEROX 3 reagent stain the sites of peroxidase activity in the granules of neutrophils, eosinophils, and monocytes. Lymphocytes, basophils, and large unstained cells contain no granules with peroxidase enzyme activity. A constant volume of the cell suspension from the peroxidase reaction chamber passes through the flowcell. The two fluids flow as independent, concentric streams (no mixing), with the ADVIA® 120 PEROX SHEATH stream encasing the sample stream. The absorbance and the forward light-scattering signatures of each blood cell are measured. The optical signals are converted to electrical pulses by photodiodes. After processing, the information is displayed in two histograms. The Perox Y histogram contains the forward-scattering data (cell size). The Perox X histogram contains the absorption data (peroxidase staining). The two histograms are combined to form the Perox cytogram from which cells are identified and counted.
Basophil/lobularity method	When the EDTA anticoagulated whole blood sample is mixed with ADVIA® 120 BASO reagent, the red blood cells are hemolyzed and the cytoplasm is stripped from all white cells except basophils. The sample is then analyzed by two-angle laser light scattering detection using a laser diode. The white cells are classified into three categories: basophils, mononuclear cells, and polymorphonuclear cells.

Note: ADVIA® 120 BASO from Siemens (Bayer/Technicon, Munich, Germany).
Abbreviation: EDTA, ethylenediaminetetraacetic acid.

expected to have a higher morbidity than the background population because these people were referred for blood testing. However, the OR calculated from this sample can be transferred to the background population of all Danes and interpreted as a relative risk if, as is often the case, the outcome is rare.[3] The disadvantages of this approach are 1) much of the information in the timing of the occurrence after the exposure is lost and 2) individuals who die or emigrate in the fixed time period after the exposure have an artificially low probability of experiencing the outcome.

- The second approach uses survival analysis, eg, Cox proportional hazard regression, to model the time until first occurrence after exposure. The advantages of this approach are 1) various follow-up periods are allowed for instead of a single one and 2) death, emigration, and other reasons for differing follow-up periods are accounted for by censoring. The disadvantages are 1) the incidence rate ratio (or hazard ratio [HR]) that is the effect measure in a Cox regression is not invariant to the prevalence of the outcome. For this reason, the result, in principle, cannot be transferred to the Danish population in general, and 2) the large amount of data in the CopDiff database will cause any test of the proportional hazard assumption to reject this with high probability. This will change the focus of the analysis toward investigating the development of the exposure effect over time. Since the timing of the exposure (blood sampling) is not a well-defined time point in the development of the disease, this time stratification seems inappropriate.

We have a slight preference for the first cohort methodology because of its epidemiological simplicity and straightforward interpretation. Only if we can attach clinical significance to the timing of the exposure, eg, if it is a diagnostic measurement of some sort, a survival analysis may be more relevant.

The two cohort design approaches are illustrated in the following data example.

Illustrative example

For this analysis we included all adults aged 18 to 80 years from the CopDiff database. From each of these 359,950 unique individuals, with at least one DIFF in the period January 1, 2001 to December 31, 2007, we randomly selected a single DIFF that contained an eosinophil count. These individuals were then categorized according to the degree of eosinophilia. Individuals with missing values for the eosinophil count (n=3,754) were excluded from the cohort. As a potential confounder, the level of C-reactive-protein (CRP), categorized as "increased" ($\geq$10 mg/L) versus "normal" ($<$10 mg/L) was also obtained from the CopDiff database. A third category was defined for those individuals for whom CRP was not measured. We computed Charlson's comorbidity index[4] from the hospital contacts recorded in the Danish National Patient Register for the 3 years before the index DIFF. Furthermore, we recorded whether another DIFF was made during the 6 months before the request and whether eosinophilia was present in this DIFF. The objective of the analysis was to investigate whether eosinophilia was associated with increased incidence of hematological malignancies (as recorded in the Danish Cancer Registry) in the period following the selected DIFF; in the following we illustrate the two approaches. Both analyses estimate the effects of eosinophilia adjusted for sex, age (quadratic), year, month, previous cancer, Charlson's comorbidity index, CRP, and previous eosinophilia.

Analysis approach A: logistic regression

The first approach analyzes the 3-year incidences of hematological malignancies in a multivariate logistic regression model. The effects of eosinophilia were estimated with OR (95% confidence interval):

- mild versus no eosinophilia: 1.36 (1.02–1.80)
- moderate versus no eosinophilia: 3.41 (1.75–6.65)
- severe versus no eosinophilia: 5.98 (3.03–11.78)

These results clearly show a trend toward higher hematological malignancy incidence with higher degree of eosinophilia. However, in a parallel analysis a similar trend could be seen for mortality. Hence, the incidence of hematological malignancy may be artificially low for the more severe eosinophilia cases, which causes the effects to be less pronounced than they should have been.

Analysis approach B: Cox proportional hazard regression

The second approach analyzes the time to the first occurrence of hematological malignancy in a multivariate Cox proportional hazard regression model, or to death or end of follow-up. The effects of eosinophilia were estimated with HR (95% confidence interval):

- mild versus no eosinophilia: 1.38 (1.09–1.75)
- moderate versus no eosinophilia: 3.11 (1.71–5.65)
- severe versus no eosinophilia: 4.88 (2.61–9.14)

This analysis also shows a clear trend: more severe eosinophilia is associated with higher incidence of hematological malignancies. Although the results from the two approaches are numerically quite similar, the two different effect measures are not comparable. However, if the event is rare and the proportional hazard assumption is true, the 3-year incidence will be similar to the hazard at 3 years, and the OR and HR will be numerically similar.

A problem with the second approach is the proportional hazard assumption. A statistical test for this assumption, eg, a likelihood ratio test for the addition of interactions of all covariates with log (time), will be overpowered. Moreover, such a test was not possible in SAS PROC PHREG

(SAS Institute Inc., Cary, NC, USA), as the estimation of a model including interactions with log (time) took too long to compute. A graphical test (as implemented in SAS PROC PHREG) produces for each covariate in the model a plot of the observed score process against several score processes simulated assuming proportional hazards. If the observed process is different from the simulated processes, the proportional hazard assumption is considered to be violated.[5] For large databases such as CopDiff, this may take a long computing time if the event of interest is not rare. For the above analysis, two such plots are shown in Figure S1. Figure S1A indicates that in relation to the mild versus no eosinophilia effect, the proportional hazard assumption holds. However, in relation to the previous DIFF effect, incidence is higher than expected in the first years after the selected DIFF (Figure S1B). Similar patterns are seen for some other covariates. The proportional hazard assumption may be handled by either estimating the baseline hazard separately, in strata spanned by categories of the violating variables, or by splitting time up into separate periods for which separate effects are calculated. To the extent this is possible given computation times, this will blur the interpretation of the effects of eosinophilia on hematological malignancies.

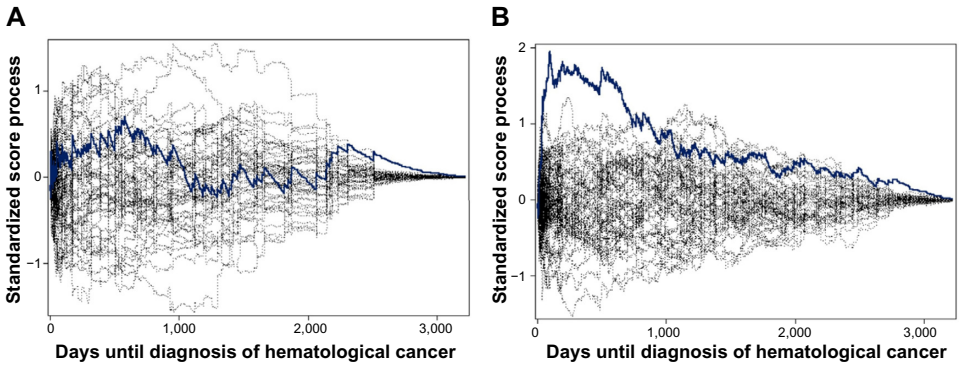

Figure S1 Observed and simulated score processes for mild eosinophilia (**A**) and previous DIFF (**B**).
Abbreviation: DIFF, differential blood cell count.

References

1. Siemens laboratory diagnostic services [webpage on the Internet]. Available from: http://www.healthcare.siemens.com/services/laboratory-diagnostics. Accessed May 18, 2014
2. Andersen CL, Siersma VD, Hasselbalch HC, et al. Eosinophilia in routine blood samples and the subsequent risk of hematological malignancies and death. *Am J Hematol*. 2013;88(10):843–847.
3. Woodward M. *Epidemiology: Study Design And Data Analysis*. 2nd ed. Chatfield C, Zidek JV, editors. Boca Rotan, FL: Chapman and Hall/CRC; 2005;124–126.
4. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Chronic Dis*. 1987;40(5):373–383.
5. Lin DY, Wei LJ, Ying Z. Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika*. 1993;80(3): 557–572.