Title: Development and validation of an algorithm to accurately identify eczema patients in primary care electronic health records from the UK

Short title: Validation of eczema in electronic health records

Authors: Abuabara K[1], Magyari AM[2], Hoffstad O[3], Jabbar-Lopez ZK[4], Smeeth L[5], Williams HC[6], Gelfand JM[3,7], Margolis DJ[3,7], Langan SM[5]

Affiliations:
1. Program in Clinical Research, Department of Dermatology, University of California San Francisco
2. Department of Health Policy & Management, UC Berkeley School of Public Health
3. Department of Biostatistics and Epidemiology, University of Pennsylvania Perelman School of Medicine
4. Unit for Population-Based Dermatology Research, St John's Institute of Dermatology
5. Faculty of Epidemiology & Population Health, London School of Hygiene and Tropical Medicine
6. Centre of Evidence-Based Dermatology, University of Nottingham
7. Department of Dermatology, University of Pennsylvania Perelman School of Medicine

Keywords: eczema, atopic dermatitis, validation, routinely collected data, prevalence, diagnosis

Conflicts: In the previous 12 months, JMG served as a consultant for Abbvie., Coherus, Janssen Biologics (formerly Centocor), Merck, Novartis Corp, Valeant, and Pfizer Inc., receiving honoraria; and receives research grants (to the Trustees of the University of Pennsylvania) from Abbvie, Eli Lilly, Janssen, Novartis Corp, Regeneron, Sanofi, and Pfizer Inc.; and received payment for continuing medical education work related to psoriasis that was supported indirectly by Lilly, and Abbvie. In addition, Dr. Gelfand is a co-patent holder of resiquimod for treatment of cutaneous T cell lymphoma. The other authors report no conflicts.

Correspondence to: Katrina.abuabara@ucsf.edu Department of Dermatology, University of California San Francisco (UCSF), 2340 Sutter Street, N421, San Francisco, CA 94115. Phone: 415-502-0433

Abbreviations used: Positive Predictive Value (PPV), CI (Confidence interval), The Health Improvement Network (THIN)

ABSTRACT (current word count: 195, limit 200)

Electronic health records offer great research scope. Undertaking eczema research using such data is challenging due to its episodic and heterogeneous nature. We sought to develop and validate a diagnostic algorithm that identifies eczema cases based on codes used for electronic records used in the UK Health Improvement Network (THIN). We found that at least one of 5 diagnosis codes plus two treatment codes for any skin-directed therapy were likely to accurately identify patients with eczema. To validate this algorithm, a questionnaire was sent to the physicians of 200 randomly selected children and adults. The primary outcome, the positive predictive value (PPV) for a physician-confirmed diagnosis of eczema, was 86% (95%CI 80-91%). Additional criteria increased the PPV up to 95% but would miss up to 89% of individuals with physician-confirmed eczema. The first and last entered diagnosis codes for individuals showed good agreement with the physician-confirmed age at onset and last disease activity; the mean difference was 0.8 years (95% CI -0.3,1.9) and -1.3 years respectively (95%CI -2.5, -0.1). A combination of diagnostic and prescription codes can be used to reliably estimate the diagnosis and duration of eczema from the THIN primary care electronic health records in the UK.

INTRODUCTION

Eczema (synonymous with atopic dermatitis or atopic eczema (Johansson *et al.*, 2004)) is one of the 50 most burdensome diseases worldwide (Vos *et al.*, 2012; Weidinger and Novak, 2016). Therefore there is great interest in understanding its causes, natural history and potential associations with comorbid conditions. Yet most studies rely on highly selected specialty clinic populations, cross-sectional studies, or self-reported data, and are prone to bias and limited generalizability (Asher *et al.*, 1995, Deckert *et al.*, 2014). Representative population-level data with validated diagnoses and longitudinal follow up are needed.

Electronic health data from primary care practices in the UK present an opportunity to directly address many of the unanswered questions about long-term outcomes in eczema in particular. They are representative of the general population, include relatively long term follow up of both children and adults, and are appropriate for the study of eczema since 97% of patients are managed by general practitioners in the UK (Emerson *et al.*, 1998; Schofield JK, 2009). However, these data were created for administrative and clinical purposes, not designed specifically for research, and it is therefore critically important that the validity of eczema diagnoses in these data sources is understood (Manuel *et al.*, 2010). Because eczema is a heterogeneous and episodic condition with non-specific terminology, there exists high potential for misclassification of diagnosis and duration of disease. There is no single diagnostic test for eczema and it can be challenging to diagnose in population-based studies due to its variability in morphology, distribution and periodicity. The diagnosis relies on clinical judgment based on a combination of history and physical examination. Previous studies using UK primary care data to

identify patients with eczema report wide variations in prevalence from 0-38% based on the coding algorithm used (Anandan *et al.*, 2009; Carey *et al.*, 2003; McKeever *et al.*, 2001; McKeever *et al.*, 2002, 2004; Punekar and Sheikh, 2009; Simpson *et al.*, 2002; Simpson *et al.*, 2009). Moreover, there is some evidence that chronic diseases, such as eczema, may be more poorly recorded over time in UK general practice data, as general practitioners are not required to enter codes on each occasion for chronic conditions (Jordan *et al.*, 2004; Khan *et al.*, 2010).

This study aimed to enhance identification of patients with eczema within electronic health records. The objectives were to develop and validate a diagnostic algorithm for eczema that identifies cases based on codes, and secondarily, to examine the agreement between physician report and codes for eczema disease onset, duration and severity.

RESULTS

<u>Algorithm development</u>

A list of potential eczema diagnosis and treatment codes were developed by employing a keyword search and examining affiliated codes (Supplemental Table 1), and the five most common and specific codes for eczema were chosen to identify those likely to have eczema: m111.00 atopic dermatitis/eczema, m1120.0 infantile eczema, m113.00 flexural eczema, m11400 allergic/intrinsic eczema, m12z100 eczema NOS. When we examined the frequency of medical codes among individual patients, we found that including 32 codes likely to be related to eczema rather than only the 5 most common codes only slightly increased the number of individuals identified, but including up to 74 possible eczema codes nearly doubled the number

4

of individuals identified (Table 1). The distribution of some codes varied between children and adults; for example, m1120.0/infantile eczema was more commonly used in children.

Despite the chronicity of eczema, any of the 5 most common diagnosis codes were rarely repeated in the database; overall, patients had a mean of 1.2 (standard deviation 0.5) codes during 5.6 years (standard deviation 8.0) of follow-up. Because eczema is by definition a chronic condition, it was important to include more than one code in our algorithm, but requiring individuals to have two or more diagnosis codes would exclude >80% of the potential eczema population. Therefore, the distribution of treatment codes was also examined. In the UK, medical record codes and treatment codes can be entered independently (i.e. a prescription code does not require an associated diagnostic code). Prescriptions, including emollient preparations, are available through the National Health Service, so we examined prescription codes for all potential relevant therapies including topical emollients, topical steroids, topical calcineurin inhibitors, topical anti-infective treatments, and systemic immunomodulatory medications (including methotrexate, azathioprine, mycophenolate, cyclosporine, or biologics) based on British National Formulary groupings, and phototherapy codes (British National Formulary, 2016; Supplemental Table 2). Since prescriptions are free of charge for children only, we stratified our analyses by age (i.e. children under 18 versus adults). We also specifically examined the use of topical steroids and topical calcineurin inhibitors (which are likely to be more specific for eczema). To ensure we captured patients with chronic eczema in our algorithm, we chose to include patients with at least one of the 5 medical codes frequently used for eczema as listed above *and* at least 2 treatment codes for any eczema-related therapy on separate dates

(at any time point relative to the eczema diagnosis, since symptoms may precede the actual diagnosis).

Physician Survey

To validate the algorithm for eczema, we surveyed the physicians of a random sample of 100 children (< 18 years of age) and 100 adults (Figure 1). The response rate was 97% overall (96% for adults and 97% for children), and there was no significant difference in response rate by age or sex. The algorithm for identifying patients with eczema performed well and there were no significant differences in codes between those with and without physician confirmed eczema (Table 2). The positive predictive value (PPV) for a single diagnostic code and at least two treatment codes was 86% overall (95%CI 80-91%); and was higher among children (90%) than adults (82%), though this difference was not statistically significant (Pearson chi2=2.76, p=0.097).

When we examined whether the use of more stringent criteria would improve the prediction of physician-confirmed eczema, we found that adding additional criteria to the algorithm had the potential to increase the PPV, but would result in smaller numbers of individuals being detected (Table 3). For example, requiring two eczema codes would increase the PPV to 91%, but would only detect 83/163 or 51% of those with physician-confirmed eczema. Similarly, requiring a dermatology consult code in addition to the eczema and prescription codes would increase the PPV to 95%, but would only detect 18/163 or 11% of those with physician-confirmed eczema. Requiring the prescriptions to be for medications more specific to eczema (i.e. topical steroids or calcineurin inhibitors) did not significantly change the PPV.

6

The average age of onset and oldest age of disease activity requiring physician contact estimated using codes from the database were similar to what physicians reported (Table 4). The mean estimated age at onset using the first diagnosis code or first treatment code were both slightly younger than the physician estimate (mean difference 0.4-0.8 years), and the level of agreement was high (76% of estimates were within one year of each other, Bland Altman plot shown in Supplemental Figure 4). The mean estimated age at last date of eczema activity using the last diagnosis code or last treatment code were both older than the physician estimate (mean difference 1.3-3.9 years), and the level of agreement was again high (79% of estimates within five years of each other, Bland Altman plots in Supplemental Figure 4). When we stratified these estimates by age comparing children under age 18 to adults we found similar results (Supplemental Table 4).

In our sample, 48 patients were reported by the physician to have had symptoms in the year prior to their last visit; 27 (56%) of whom were assessed as having mild disease and 19 (40%) of who were assessed as having moderate disease based on the severity descriptions in the National Institute for Health and Care Excellence (NICE) guidelines. Patients with moderate disease had more treatment codes during that year than patients with mild disease (median 5 versus 2, p-value for two-sample Wilcoxon rank sum test =0.887). None were reported to have severe disease, limiting our ability to draw any conclusions about the validity of medical record codes to predict disease severity.

Finally, we assessed whether physicians would be able to adequately respond to the UK Working Party criteria (originally designed for in-person assessment), enabling us to compare a set of well-validated criteria for use in large epidemiologic studies to our outcomes in routinely collected electronic health data. For each question, we gave physicians the option of choosing 'don't know'. The high number of uncertain responses resulted in poor ability to discriminate between those with and without eczema (Table 5). We found that only 52 (32%) of those with physician-confirmed eczema in our sample met the criteria (an itchy skin condition plus at least 3 of history of flexural involvement, history of asthma/hay fever, history of generalized dry skin, onset of rash under age 2, and visible flexural dermatitis).

DISCUSSION

Interpretation of main findings

Patients with eczema were accurately identified if they had at least one eczema diagnostic code and at least two prescription codes for eczema-related treatments in a large electronic medical record database representative of the general population in the UK. The positive predictive value, or probability that individuals identified by our algorithm truly have the disease as determined by their doctor, was 86%, which is similar to the PPV of coding algorithms for other chronic diseases in routinely collected data (Khan *et al.*, 2010). The PPV was higher in children, but the algorithm still performed well to identify adults with eczema.

This study indicates that the types, number, and frequency of codes used to identify eczema patients in routinely collected data are important because small differences have the potential to

cause substantial misclassification. After examining the distribution of all of codes potentially related to eczema, we chose to use the 5 most common eczema codes in addition to treatment codes for the primary algorithm. As shown in Table 1, expanding the definition from 5 to 32 codes (likely related to eczema but rarely used) would have only increased the proportion of the population identified from 13 to 14%, so we opted for the more parsimonious algorithm. In contrast, using a single code to define eczema, for example AD/Eczema (M111.00), would identify far fewer individuals (only 6% of the population). Although it was impractical and prohibitively costly to sample enough physicians to calculate the sensitivity, specificity and predictive value of each of these variations, we present the proportion of patients identified by each set of codes to illustrate the potential magnitude of misclassification. We were able to calculate *post-hoc* changes in the PPV caused by adding criteria to our algorithm. Inclusion of a second diagnosis code, allergy code, or consult code all increased the PPV, but would identify far fewer patients. The ideal balance between these factors depends on the research question. For example, an algorithm with a very high PPV that captures only a fraction of those with disease may be acceptable for a case-control study. On the other hand, the ideal algorithm for a prevalence study would aim to assess the total population burden accurately and may include more mild or marginal cases.

Because eczema is a chronic condition, we explored the possibility of using codes from more than one time point to identify patients. In the UK, providers are not required to re-enter codes for chronic conditions, and only 36% of individuals had more than 1 eczema diagnosis code. Treatment codes, which can be entered independently from diagnostic codes, were used more frequently, and were therefore included in the algorithm. When selecting the treatments, we

opted for an inclusive approach and used all potential eczema-related treatments, even

emollients, as listed under British National Formulary categories. This approach may include

treatments not specifically for eczema, so we examined the performance of a more limited

definition of treatments (only topical steroids or topical calcineurin inhibitors), and found it did

not change the PPV (Table 3). Of note, 22% of individuals with one of the 5 most common

medical codes never received any treatment codes. Our algorithm excluded these patients, some

of whom may have had mild untreated disease.

Because we randomly selected individuals with eczema diagnoses at any time point, only a

fraction had disease activity during the year prior to their last visit, resulting in too few numbers

to meaningfully assess the validity of codes relative to disease severity. Additional research is

necessary to validate whether codes can be used to ascertain severity and disease flares in

routinely collected data.

Comparability to other studies

Three other studies attempted to validate routinely collected data for identifying individuals with

eczema. Two examined the use of medications alone and found they had poor discriminatory

power to identify patients with eczema in the Netherlands and Sweden (Mulder *et al.*, 2016;

Ortqvist *et al.*, 2013). The third compared ICD-9 codes from a tertiary care population in the US

with Hanifin & Rajka and UK Working Party (UKWP) criteria found in the medical record and

found poor overlap (Hsu *et al.*, 2016), possibly due to the lack of standardized recording of

specific diagnostic features in the medical record. We assessed whether it was possible to

compare our results to the UK Working Party diagnostic criteria, which have been used for

epidemiological studies in multiple international settings, but were developed for in-person assessment (Brenninkmeijer *et al.*, 2008; Williams *et al.*, 1994). Because physicians responded, "don't know" to so many of the UK Working Party questions in our survey, we were unable to make meaningful comparisons. We hypothesize the high rates of uncertainty were because there was not enough data in the medical record to enable physicians to answer all of the required questions, and therefore caution against using these as a gold standard from medical record review when they were not systematically assessed. It is also possible that those deemed to have eczema by their physician simply would not fulfill the criteria if they had been ascertained fully, and further specially designed studies are needed to test this notion.

Strengths and weaknesses

Strengths of our study include the use of diagnosis and treatment codes, stratified sampling among children and adults, a large representative database with longitudinal follow up, and physician-confirmation of disease as the gold standard. We sampled general practice physicians rather than dermatologists because 97% of patients with eczema are managed by general practitioners in the UK, and sampling specialists would have limited the generalizability of the results (Emerson *et al.*, 1998; Schofield JK, 2009).

Ideally, patients would have been assessed in person to confirm their diagnoses. Because this was not possible through the Additional Information Services in THIN, we queried their physicians instead. The physicians were asked to assess the patient based on their recall and review of the medical record. This approach was chosen over a medical record review because it

allowed for direct assessment as to whether the physician really believed the patient had eczema (regardless of coding).

Our results are only directly generalizable to The Health Improvement Network, though the algorithm is likely to perform similarly in the other UK primary care databases which have substantial overlap (the Clinical Practice Research Datalink https://www.cprd.com/, and other UK primary care data sources including QResearch http://www.qresearch.org/). The PPV is likely to differ in settings where the prevalence of eczema varies and where the data structure and incentives for coding differ. Nonetheless, our results highlight the potential biases that should be considered when selecting combinations of codes to identify eczema patients in any setting.

Implications for future research

Validation studies that ensure patients are accurately identified are a high priority to enable the use of increasingly available and robust sources of routinely collected electronic health data (De Coster *et al.*, 2006). This study showed that eczema patients can be accurately identified in the UK Health Improvement Network, and that changes in the number, type or frequency of codes used could result in large differences in the number of patients identified. We highlight factors to consider when examining the frequency and distribution of diagnostic and treatment codes in any electronic medical record database, which are important for researchers to avoid misclassification bias.

METHODS

Study design

Our study consisted of two parts: a longitudinal cohort study to develop a diagnostic algorithm, and a physician survey to validate it. We followed guidelines for reporting of validation studies and reporting of studies conducted using observational routinely collected health data (Benchimol *et al.*, 2011; Benchimol *et al.*, 2016).

Participants/Data source

The Health Improvement Network (THIN) is a database comprising the electronic health records of people registered with participating general practices. THIN is broadly representative of the general UK population in terms of age, sex, ethnicity, and geography and is one of three major UK primary care databases (Shephard *et al.*, 2011). We chose this data source because it is one of the world's largest sources of anonymized longitudinal data from primary care practices with over 85 million patient-years of follow up, and because we had institutional access and experience using the data (Margolis *et al.*, 2007; Margolis *et al.*, 2008; Ogdie *et al.*, 2015; Seminara *et al.*, 2011). Previous validation studies have shown that the recording is highly accurate and nearly complete, and THIN has been used to study multiple chronic conditions. Participating practices are remunerated for recording data on clinical diagnoses, test results, prescriptions, and referral data via the Read/OXMIS (Oxford Medical Information System) coding framework, which is based on the International Classification of Diseases (ICD) coding system. The raw data are updated monthly and undergo extensive quality control and validity checks by a centralized research team before release. Practices may choose to participate in the

Additional Information Services Program, which administers surveys to consenting physician practices. Approximately 60% of all THIN practices actively participated in this program when our survey was administered in October 2015.

Algorithm development

A list of potential eczema diagnosis and treatment codes were developed by employing a keyword search and examining affiliated codes (Supplemental Table 1). The distribution of codes was examined, and in consultation with a panel of experts on eczema epidemiology and use of routinely collected data (HCW, DM, LM, SML, KA) a parsimonious algorithm was developed to identify patients most likely to have eczema.

Physician Survey

The survey was sent to the physicians of a random sample of 100 children (<18 years of age) and 100 adults with acceptable records who were alive and currently enrolled in practices participating in the Additional Information Services (Figure 1). The primary outcome was the positive predictive value (PPV), or probability that subjects identified by the algorithm truly have the disease, as this measure is the most relevant for avoiding misclassification bias in subsequent studies of eczema (Choi, 1992). Assuming a physician response rate of 90% (based on prior studies using physician confirmation of chronic disease in routinely collected data (Khan *et al.*, 2010; Seminara *et al.*, 2011)), a sample of 200 patients should have enabled us to obtain a 95% confidence interval of 0.85-0.94 around an *a priori* estimated PPV of 0.90. Given funding constraints we chose to sample only patients with codes suggestive of eczema. Sampling

additional subjects without eczema codes would have enabled us to also calculate sensitivity and specificity of the algorithm.

A standardized letter was sent to each practice requesting completion of a 1-page survey (supplemental Figures 2-3), and physicians received monthly reminders for completion and compensation for their time. If the diagnosis of eczema was confirmed, we then asked the physician to (1) provide a global assessment of average eczema severity over the past 12 months, (2) confirm the age at eczema onset, and (3) confirm whether the patient still has active eczema or whether the patient's eczema is in remission. Although many eczema-specific severity scales have been developed and validated for assessment of patient outcomes in clinical trials, few are designed to address long-term severity (Schmitt *et al.*, 2007). Therefore, to asses severity, we used descriptions of mild, moderate, and severe disease from the UK National Institute for Health and Care Excellence (NICE) guidelines for management of eczema (Excellence, 2007). Finally, to determine whether our results could be compared to another widely used definition of eczema in large epidemiologic studies, the survey included the UK Working Party refinement of Hanifin and Rajka's diagnostic criteria questions (Brenninkmeijer *et al.*, 2008; Williams *et al.*, 1994).

Eczema is a clinical diagnosis, and biopsy and laboratory tests are non-specific, therefore we relied on the physician's confirmation of the diagnosis as the gold standard. This approach is consistent with other validation studies of chronic conditions in medical record databases in UK primary care databases (Ogdie *et al.*, 2014; Seminara *et al.*, 2011; Soriano *et al.*, 2001).

Physicians were asked to fill out the survey based on their knowledge of the patient and review of his or her medical record.

Analysis

For the 200 patients whose physicians were surveyed, differences in codes between those with and without physician-confirmed eczema were examined and the PPV of our algorithm for identifying eczema patients was calculated. The PPVs of alternative algorithms with additional criteria for identifying patients with eczema were also calculated. Next, the age of disease onset and "remission" reported in the physician survey were compared to dates calculated from the database using the first and last eczema diagnosis and prescription codes. Agreement was assessed using Bland Altman plots (Bland and Altman, 1986). All analyses were stratified by age (i.e. children under 18 *vs* adults). Analyses were performed using Stata (Version 14, Stata Corporation, College Station, Tx).

Ethics

Approval was obtained from the Scientific Research Council of THIN and the University of Pennsylvania IRB.

REFERENCES

Anandan C, Gupta R, Simpson CR*, et al.* (2009) Epidemiology and disease burden from allergic disease in Scotland: analyses of national databases. *J R Soc Med* 102:431-42.

Asher MI, Keil U, Anderson HR*, et al.* (1995) International Study of Asthma and Allergies in Childhood (ISAAC): rationale and methods. *Eur Respir J* 8:483-91.

Benchimol EI, Manuel DG, To T*, et al.* (2011) Development and use of reporting guidelines for assessing the quality of validation studies of health administrative data. *J Clin Epidemiol* 64:821-9.

Benchimol EI, Smeeth L, Guttmann A*, et al.* (2016) [The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement]. *Zeitschrift fur Evidenz, Fortbildung und Qualitat im Gesundheitswesen* 115-116:33-48.

Bland JM, Altman DG (1986) Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1:307-10.

Brenninkmeijer EE, Schram ME, Leeflang MM*, et al.* (2008) Diagnostic criteria for atopic dermatitis: a systematic review. *Br J Dermatol* 158:754-65.

British National Formulary <https://www.bnf.org/products/bnf-online/ > Accessed Dec 30 2016.

Carey IM, Cook DG, De Wilde S, *et al.* (2003) Implications of the problem orientated medical record (POMR) for research using electronic GP databases: a comparison of the Doctors Independent Network Database (DIN) and the General Practice Research Database (GPRD). *BMC family practice* 4:14.

Choi BC (1992) Sensitivity and specificity of a single diagnostic test in the presence of work-up bias. *J Clin Epidemiol* 45:581-6.

De Coster C, Quan H, Finlayson A, *et al.* (2006) Identifying priorities in methodological research using ICD-9-CM and ICD-10 administrative data: report from an international consortium. *BMC Health Serv Res* 6:77.

Deckert S, Kopkow C, Schmitt J (2014) Nonallergic comorbidities of atopic eczema: an overview of systematic reviews. *Allergy* 69:37-45.

Emerson RM, Williams HC, Allen BR (1998) Severity distribution of atopic dermatitis in the community and its relationship to secondary referral. *Br J Dermatol* 139:73-6.

Hsu DY, Dalal P, Sable KA, *et al.* (2016) Validation of international classification of disease ninth revision codes for atopic dermatitis. *Allergy*.

Johansson SG, Bieber T, Dahl R, *et al.* (2004) Revised nomenclature for allergy for global use: Report of the Nomenclature Review Committee of the World Allergy Organization, October 2003. *J Allergy Clin Immunol* 113:832-6.

Jordan K, Porcheret M, Croft P (2004) Quality of morbidity coding in general practice computerized medical records: a systematic review. *Fam Pract* 21:396-412.

Khan NF, Harrison SE, Rose PW (2010) Validity of diagnostic coding within the General Practice Research Database: a systematic review. *Br J Gen Pract* 60:e128-36.

Manuel DG, Rosella LC, Stukel TA (2010) Importance of accurately identifying disease in studies using electronic health records. *BMJ* 341:c4226.

Margolis DJ, Hoffstad O, Bilker W (2007) Association or lack of association between tetracycline class antibiotics used for acne vulgaris and lupus erythematosus. *Br J Dermatol* 157:540-6.

Margolis DJ, Hoffstad O, Strom BL (2008) Association between serious ischemic cardiac outcomes and medications used to treat diabetes. *Pharmacoepidemiol Drug Saf* 17:753-9.

McKeever TM, Lewis SA, Smith C, *et al.* (2001) Siblings, multiple births, and the incidence of allergic disease: a birth cohort study using the West Midlands general practice research database. *Thorax* 56:758-62.

McKeever TM, Lewis SA, Smith C*, et al.* (2002) Mode of delivery and risk of developing allergic disease. *J Allergy Clin Immunol* 109:800-2.

McKeever TM, Lewis SA, Smith C*, et al.* (2004) Vaccination and allergic disease: a birth cohort study. *American journal of public health* 94:985-9.

Mulder B, Groenhof F, Kocabas LI*, et al.* (2016) Identification of Dutch children diagnosed with atopic diseases using prescription data: a validation study. *Eur J Clin Pharmacol* 72:73-82.

National Institute for Health and Care Excellence (2007) NICE guidelines [CG57] Atopic eczema in children: Management of atopic eczema in children from birth up to the age of 12 years. <https://www.nice.org.uk/guidance/cg57/chapter/Key-priorities-for-implementation> Accessed 30 Decemeber, 2016.

Ogdie A, Alehashemi S, Love TJ*, et al.* (2014) Validity of psoriatic arthritis and capture of disease modifying antirheumatic drugs in the health improvement network. *Pharmacoepidemiology and drug safety* 23:918-22.

Ogdie A, Yu Y, Haynes K*, et al.* (2015) Risk of major cardiovascular events in patients with psoriatic arthritis, psoriasis and rheumatoid arthritis: a population-based cohort study. *Annals of the rheumatic diseases* 74:326-32.

Ortqvist AK, Lundholm C, Wettermark B, *et al.* (2013) Validation of asthma and eczema in population-based Swedish drug and patient registers. *Pharmacoepidemiol Drug Saf* 22:850-60.

Punekar YS, Sheikh A (2009) Establishing the incidence and prevalence of clinician-diagnosed allergic conditions in children and adolescents using routinely collected data from general practices. *Clin Exp Allergy* 39:1209-16.

Schmitt J, Langan S, Williams HC, *et al.* (2007) What are the best outcome measurements for atopic eczema? A systematic review. *The Journal of allergy and clinical immunology* 120:1389-98.

Schofield JK WH (2009) Skin Conditions in the UK: A Health Care Needs Assessment. . University of Nottingham.

Seminara NM, Abuabara K, Shin DB, *et al.* (2011) Validity of The Health Improvement Network (THIN) for the study of psoriasis. *Br J Dermatol* 164:602-9.

Shephard E, Stapley S, Hamilton W (2011) The use of electronic databases in primary care research. *Fam Pract* 28:352-4.

Simpson CR, Anderson WJ, Helms PJ, *et al.* (2002) Coincidence of immune-mediated diseases driven by Th1 and Th2 subsets suggests a common aetiology. A population-based study using computerized general practice data. *Clin Exp Allergy* 32:37-42.

Simpson CR, Newton J, Hippisley-Cox J*, et al.* (2009) Trends in the epidemiology and prescribing of medication for eczema in England. *J R Soc Med* 102:108-17.

Soriano JB, Maier WC, Visick G*, et al.* (2001) Validation of general practitioner-diagnosed COPD in the UK General Practice Research Database. *Eur J Epidemiol* 17:1075-80.

Vos T, Flaxman AD, Naghavi M*, et al.* (2012) Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet* 380:2163-96.

Weidinger S, Novak N (2016) Atopic dermatitis. *Lancet* 387:1109-22.

Williams HC, Burney PG, Hay RJ*, et al.* (1994) The U.K. Working Party's Diagnostic Criteria for Atopic Dermatitis. I. Derivation of a minimum set of discriminators for atopic dermatitis. *Br J Dermatol* 131:383-96.

Figure 1. Flow chart showing the sampling of patients from THIN and resulting classification

Any of 5* eczema med codes +
2 Rx codes on different days
N=1,280,906

Transferred out, died, other (40%)
Unacceptable record (4%)
Practice doesn't participate in AIS (66%)

Number of eligible cases
available for validation
215,320

Random Sample
100 children
100 adults

7 Cases (4 adults and 3 children)
with no response

193 surveys (96 children and 97
adults) available for analysis
189 surveys with data on eczema

163 (86%)
Physician confirmed
diagnoses

26 (14%)
Number of cases not
validated

Table 1. Distribution of codes in the entire THIN database

	Total	Children (ages 0-17)	Adults (ages 18+)
	N = 9,775,618	**N = 1,404,158**	**N = 8,371,460**
Diagnosis codes			
Proportion of the total population with each of the following medical codes			
Atopic Dermatitis/ Eczema M111.00	6%	13%	5%
Infantile Eczema M112.00	1%	7%	0%
Flexural Eczema M113.00	1%	2%	0%
Allergic/Intrinsic Eczema M114.00	0%	0%	0%
Eczema NOS M12z100	6%	8%	6%
Proportion of the total population with at least one of the 5 codes listed above	13%	23%	11%
Proportion of the total population with at least 2 of the 5 codes listed above	4%	10%	4%
Proportion of the total population with at least one of the 32 likely eczema codes*	14%	25%	13%
Proportion of the total population with at least one of the 74 possible eczema codes*	29%	47%	26%

Notes: *See Supplemental Table 1. Ages as of Jan 2013; note that among adults, codes may have occurred before age 18.

Table 2. Survey sample characteristics

Survey responses	Total*	Confirmed eczema*	No eczema*	Chi² or Fisher's p-value
	200	163 (81.5%)	26 (13%)	
Diagnosis codes				
Atopic Dermatitis/ Eczema M111.00	116 (58%)	98 (60%)	13 (50%)	0.330
Infantile Eczema M112.00	30 (15%)	24 (15%)	4 (15%)	0.930
Flexural Eczema M113.00	16 (8%)	13 (8%)	2 (8%)	0.960
Allergic/Intrinsic Eczema M114.00	3 (2%)	3 (2%)	0 (0%)	0.486
Eczema NOS M12z100	86 (43%)	73 (45%)	9 (34%)	0.331
Mean number (SD) of the 5 eczema codes listed above	1.3 (0.5)	1.3 (0.6)	1.1 (0.3)	0.051
Mean number (SD) of 32 likely eczema codes*	2.6 (2.9)	2.8 (3.1)	1.7 (1.2)	0.070
Mean number (SD) of 74 possible eczema codes*	4.0 (3.8)	4.1 (3.9)	3.2 (3.3)	0.271
Prescription codes				
Mean number (SD) of prescriptions for any eczema-related therapy*	16.3 (24.5)	17.5 (26)	11.1 (15.0)	0.226
Mean number (SD) of topical steroid or calcineurin inhibitor prescriptions*	9 (15)	6.6 (8.9)	6.5 (7.9)	0.953
Mean number (SD) of systemic medication codes*	0.4 (5.8)	0.5 (6.4)	0 (0)	0.691
Other				
Mean number (SD) of exclusionary diagnostic codes*	0.4 (1.7)	0.3 (0.7)	0.5 (1.6)	0.281
Total (%) with at least one exclusionary condition*	29 (15%)	25 (15.3%)	3 (11.5%)	
Mean number (SD) of diagnostic procedure (biopsy or patch testing) codes*	0	0	0	N/A
Mean number (SD) of dermatology consultation codes*	0.2 (1.1)	0.3 (1.6)	0.0 (0.2)	0.308
Total (%) with at least one dermatology consult code	19 (10%)	18 (11%)	1 (4%)	
History of atopy, N (%)**	64 (39%)	56 (41%)	6 (24%)	0.110
Male N (%)	100 (50%)	86 (53%)	9 (35%)	0.086

Notes: *Columns do not sum to 200 because of missing values (7 unreturned surveys and 4 returned surveys missing a response to the eczema question). See Supplemental Table 1 for specific codes. Any eczema related therapy includes topical skin preparations, topical steroids, topical calcineurin inhibitors, topical anti-infective treatments, systemic immunomodulatory medications (including methotrexate, azathioprine, mycophenolate, cyclosporine, or biologics), and phototherapy.
**Per physician response on survey; defined as a history of "other atopic disease (e.g. asthma or allergic rhinitis) for adults OR a family history of atopic disease in a first degree relative if aged under 4 years"

Table 3. Positive Predictive Value of Coding Algorithms

	True Positives/ All Positives	% of patients with confirmed eczema identified	All		Children (ages 0-17)		Adults (ages 18+)	
Baseline algorithm			PPV (%)	95%CI	PPV (%)	95%CI	PPV (%)	95%CI
One of 5 eczema codes + at least 2 treatment codes on separate dates (survey selection criteria)	163/189	N/A	86%	(80-91%)	90%	(83-96%)	82%	(73-89%)
Alternative algorithms*								
Baseline algorithm; at least one treatment is a topical steroid/TCI code	157/183	96%	86%	(80-91%)	90%	(81-95%)	82%	(73-89%)
Baseline algorithm; at least one treatment is a topical steroid/TCI code either 3 months prior or up to 1 year after the eczema code	81/92	50%	88%	(80-94%)	92%	(80-98%)	84%	(70-94%)
Baseline algorithm; at least two treatments are topical steroid/TCI codes	133/153	82%	87%	(81-92%)	91%	(82-97%)	84%	(74-91%)
Baseline algorithm + an additional eczema code (2 eczema codes total)	83/91	51%	91%	(83-96%)	94%	(82-99%)	88%	(74-96%)
Baseline algorithm + an additional eczema code (2 eczema codes total); at least one treatment is a topical steroid/TCI code	82/90	50%	91%	(83-96%)	94%	(83-99%)	88%	(74-96%)
Baseline algorithm + an additional eczema code (2 eczema codes total); at least two treatments are topical steroid/TCI code	133/153	82%	87%	(81%-92%)	91%	(82%-97%)	84%	(74%-91%)
Baseline algorithm + no exclusionary condition code	138/161	85%	86%	(79-91%)	89%	(81-95%)	82%	(71-90%)
Baseline algorithm + asthma or rhinitis code	52/56	32%	93%	(83-98%)	95%	(76-100%)	91%	(77-98%)
Baseline algorithm + dermatology consult code	18/19	11%	95%	(74-100%)	100%	(54-100%)	92%	(64-100%)

Notes: *See Supplemental Tables 2 and 3 for lists of codes. TCI= topical calcineurin inhibitor

Table 4. Age at diagnosis or at last disease activity requiring contact with the physician

	Distribution of estimates by source		Difference between physician estimate from survey and database	
	Mean	95% CI	Mean	95% CI
Age in years at diagnosis (N=160)				
Physician survey	17.9	(14.3, 21.4)	N/A	N/A
Database				
First diagnosis code*	17.1	(13.5, 20.6)	0.8	(-0.3, 1.9)
First prescription for any eczema treatment**	17.4	(13.9, 21.0)	0.4	(-0.8, 1.7)
If no symptoms in the year prior to the last visit date, age at last disease activity (N=53)				
Physician survey	20.7	(14.3, 27.2)	N/A	N/A
Database				
Last diagnosis code*	22.0	(15.6, 28.5)	-1.3	(-2.5, -0.1)
Last prescription for any eczema treatment**	24.6	(14.3, 27.2)	-3.9	(-5.3, -2.4)

Notes: *Any of the 5 most commonly used codes (Atopic Dermatitis/ Eczema M111.00, Infantile Eczema M112.00, Flexural Eczema M113.00, Allergic/Intrinsic Eczema M114.00, Eczema NOS M12z100). **See Supplemental Table 2.

Table 5. Results of comparison of survey results to UK Working Party Criteria Questions

	N (%) with physician-confirmed eczema that said yes	N (%) with physician-confirmed eczema that said no	Unknown/ missing (% among those who returned a survey)
Has the patient had an itchy skin condition?	144 (88%)	8 (5%)	11 (7%)
Plus, at least 3 of the following:	52 (32%)	1 (4%)	1 (4%)
Does the patient have a history of generally dry skin?	64 (39%)	38 (23%)	61 (37%)
Has the patient ever had *visible* flexural dermatitis?	75 (46%)	49 30%)	39 (24%)
Does the patient have a *history of* flexural dermatitis?	56 (34%)	66 (40%)	41 (25%)
Does the patient have a history of other atopic disease?*	56 (34%)	81 (50%)	26 (16%)
Onset under age 2?	65 (40%)	95 (58%)	3 (1%)

Notes: *e.g. asthma or allergic rhinitis or in a first degree relative if under age 4